

On Capacity of Line Networks

Urs Niesen, Christina Fragouli, and Daniela Tuninetti

Abstract

We consider communication through a cascade of discrete memoryless channels. The source and destination node of this cascade are allowed to use coding schemes of arbitrary complexity, but the intermediate relay nodes are restricted to process only blocks of a fixed length. We investigate how the processing at the relays must be chosen in order to maximize the capacity of the cascade, that is, the maximum reliable end-to-end rate between the source and the destination. For infinite cascades with fixed intermediate processing length, we prove that the intermediate processing at the relays can be chosen to be identical without loss of optimality. In this case, we show that the capacity of the cascade coincides with the rate of the best zero error code of length equal to the blocklength of the intermediate processing. We further show that for fixed and identical intermediate processing at all relays, the limiting value of the end-to-end rate is achieved exponentially fast in the length of the cascade. Finally, we characterize how the blocklength of the intermediate processing must scale with the length of the cascade to guarantee a constant end-to-end rate. We prove that it is sufficient that the blocklength scales logarithmically with the network length in order to achieve any rates above the zero error capacity. We also show that in many cases of interest logarithmic grow is also necessary.

This work was supported in part by FNS under award No. 200021-103836/1.

U. Niesen is with MIT, Email: uniesen@mit.edu, C. Fragouli is with EPFL, Email: christina.fragouli@epfl.ch, and D. Tuninetti is with UIC, Email: daniela@ece.uic.edu.

I. INTRODUCTION

Communication systems today are organized in large scale networks, with Internet the most conspicuous example, where information needs to traverse multiple hops to reach a destination. Each of the hops may introduce errors, that become more and more pronounced as the size of the network grows. Two main approaches are used for error correction: Automatic Repeat reQuest (ARQ) schemes that achieve high reliability at the cost of delay; and packet-level Forward Error Correction (FEC) schemes, that reduce delay at the cost of reliability. The FEC schemes employed today are end-to-end: packets are encoded at the source and decoded at the destination, while intermediate nodes are only allowed to replicate and forward packets. This end-to-end approach can lead to a significant waste of resources, that is becoming less and less acceptable as multimedia applications become more popular and bandwidth demanding.

From a theoretical point of view, it is well known that decoding and re-encoding the information sent by the source (using possibly unbounded complexity) at all intermediate nodes achieves the “min-cut capacity”, as is rigorously described in [1]. Such a scheme imposes heavy computational requirements on the intermediate nodes of the network, especially since these nodes typically need to accommodate a large number of traffic connections. Moreover, it incurs large delay, that is prohibitive for real-time applications.

Recently it was demonstrated that even for lossless links, allowing intermediate nodes to process information can increase the achievable rate in a multicasting scenario [2], [3]. The proposed approach termed “network coding” requires intermediate nodes to perform linear combinations over a finite field. The complexity of the computations is hence proportional to the size of the finite field, which in turn is bounded as a function of the number of receivers [4], [5], [6]. The interesting observation is that allowing intermediate nodes to perform *finite complexity* processing may not only increase the achievable end-to-end rates, but actually achieve the min-cut capacity of the network. Moreover, the emergence of network coding helped to realize that intermediate node processing is plausible and compelling for new protocols, designed for example for overlay networks.

Motivated by these observations, in this paper we investigate what benefits finite complexity processing at intermediate nodes may offer, using information-theoretic tools. We restrict our attention to unicast communication, i.e., a single source-destination pair. In fact, today in the

Internet, almost all traffic, including multicasting, is implemented via multiple unicast sessions.

We consider a communication network where a source node transmits information to a destination node along a path that comprises L consecutive links of the network. We assume that each link corresponds to a Discrete Memoryless Channel (DMC). In other words, we model the communication path between the source and the destination as a line network consisting of L cascaded identical DMCs. This model captures both physical layer and application layer communication.

To measure complexity, we allow each intermediate node to process blocks of N symbols, and use N as complexity measure. This is a reasonable definition of complexity as it allows to bound not only processing complexity, but also delay, and memory requirements at intermediate nodes. Moreover, it is well suited to an environment where information is transmitted in packets. We allow the source and the destination to possibly code and decode over an unbounded number of length- N blocks. This is also not unrealistic, since the source and the destination have a strong incentive to devote resources towards their communication.

We are interested in the maximum information-theoretic rate that the source can reliably convey to the destination as a function of the blocklength N and the network length L . The main contributions of this work, summarized in more detail in Section II-B, are:

- We show that, if the network length increases ($L \rightarrow \infty$) but the blocklength N is fixed, the *optimal processing is identical at each relay and corresponds to a zero error code*. Thus, the capacity of the cascade coincides with the end-to-end zero error capacity. The zero error capacity is the maximum rate at which we can communicate over a channel, with zero probability of error [7]. An intuitive interpretation of this result is that, as $L \rightarrow \infty$, the zero error capacity is the only part of the transmitted information rate that we may hope to preserve.
- This limiting result, apart from its theoretical interest, applies to large networks where a packet needs to traverse a large, but not infinite, number of relays and the relays use the same fixed processing. For example, in wireless ad-hoc networks with n nodes, the average path length scales as $L = \sqrt{n}$, and it is a standard assumption that the nodes are simple identical devices [8]. In the case of identical processing at each relay, we compute the limiting end-to-end rate as $L \rightarrow \infty$ and we show that the rate of convergence to the limiting end-to-end rate is exponential in the number of cascaded channels.

- We also examine how fast the blocklength N needs to grow with network length L in order to achieve a constant fraction of the min-cut (as opposed to the zero error) capacity as $L \rightarrow \infty$. We show that logarithmic growth is sufficient to achieve any rates above the zero error capacity and it is also necessary in many cases of interest.

The capacity of a line network of identical DMCs *with identical finite complexity* intermediate processing has been investigated in [9], where lower bounds on the capacity for large N are given based on error-exponent and worst-channel case arguments. A similarly formulated problem was also announced to be under examination, however for very small values of N , in [10]. Cascades of DMCs *without processing* at the intermediate nodes have been considered also in [11], [12], [13]. The work in [11] gives an expression for the capacity of a cascade of identical channels in terms of the eigen-decomposition of the channel transition matrix, which is assumed diagonalizable and non-singular. The work in [12] looks at the capacity and at the error probability of a cascade of identical binary channels, not necessarily symmetric. The work in [13] considers the optimal ordering of different binary channels such that the capacity of the overall cascade is maximized. FEC schemes that employ intermediate processing are proposed in [14], [15]. However, these schemes are still designed assuming $N \rightarrow \infty$.

The paper is organized as follows. Section II formally introduces the network model and briefly summarizes the main results of the paper. Section III proves some basic properties of the optimal intermediate processing. Section IV calculates the capacity of an infinite cascade of identical channels without intermediate processing. Section V identifies the optimal finite length intermediate processings for an infinite cascade and establishes connections with the zero error capacity and common information. Section VI derives upper and lower bounds on the capacity. Section VII determines how the length of the processing must scale with the length of the cascade in order to achieve rate above the zero error capacity. Finally, Section VIII concludes the paper and briefly discusses open problems and future work directions.

II. NETWORK MODEL AND MAIN RESULTS

A. Network Model, Notation and Basic Concepts

We consider line networks with $L - 1$ relays as depicted in Figure 1. The source A_0 sends information to the destination A_L via relays $\{A_i\}_{i=1}^{L-1}$. Each link corresponds to the same DMC with finite input alphabet \mathcal{X} , finite output alphabet \mathcal{Y} , and arbitrary transition probability matrix

V . We impose that all the DMCs in the cascade are the same. This setup can be generalized by connecting the relays through different DMCs. Some of our results directly extend to that case, and we will point this out throughout the paper.

We will use the following notational conventions. We denote by $\mathcal{S}_{n,m}$ the set of stochastic matrices of dimension $n \times m$. We use boldface to indicate matrices or vectors. In the rest of the work logarithms are with respect to the natural basis, with the exception of plots, where capacities are in bits per channel use.

We restrict the relays $\{A_i\}_{i=1}^{L-1}$ to perform operations from blocks of N symbols in \mathcal{Y} to blocks of N symbols in \mathcal{X} in a memoryless way across blocks. Using N times the channel V between A_i and A_{i+1} , amounts to connecting A_i and A_{i+1} through an equivalent DMC with input alphabet \mathcal{X}^N , output alphabet \mathcal{Y}^N , and transition probability matrix

$$\mathbf{W} \triangleq \mathbf{V}^{\otimes N},$$

where \otimes denotes the Kronecker product. For the node A_i , we denote by $\mathbf{X}_i \in \mathcal{X}^N$ what the relay sends and with $\mathbf{Y}_i \in \mathcal{Y}^N$ what the relay receives. The output \mathbf{X}_i is then a (not necessarily deterministic) function of \mathbf{Y}_i . This function can be described by a transition probability matrix $\mathbf{M}_i \in \mathcal{S}_{|\mathcal{Y}^N|, |\mathcal{X}^N|}$ specifying for each realization \mathbf{x} of \mathbf{X}_i and \mathbf{y} of \mathbf{Y}_i the probability $\Pr[\mathbf{X}_i = \mathbf{x} | \mathbf{Y}_i = \mathbf{y}]$.

We allow the source A_0 and the destination A_L to perform coding and decoding of arbitrary complexity, across a possibly infinite number of symbols in \mathcal{X}^N and \mathcal{Y}^N .

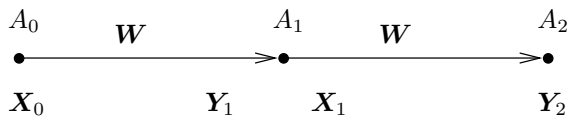


Fig. 1. A line network with two channels and one relay ($L = 2$).

We are interested in identifying the set of processings $\{\mathbf{M}_i\}_{i=1}^{L-1}$ performed at the relays that maximize the achievable rate between the source and the destination. This is exactly the capacity of the overall channel

$$\mathbf{W}_{\text{eq}}(\{\mathbf{M}_i\}) \triangleq \mathbf{W} \prod_{i=1}^{L-1} (\mathbf{M}_i \mathbf{W}) \quad (1)$$

that includes the intermediate processing $\{\mathbf{M}_i\}_{i=1}^{L-1}$ as part of the channel transition probability matrix. Our goal is to determine

$$C_{N,L}(\mathbf{V}) \triangleq \max_{\{\mathbf{M}_i\}_{i=1}^{L-1}} \frac{1}{N} C(\mathbf{W}_{\text{eq}}(\{\mathbf{M}_i\})),$$

the capacity of the channel \mathbf{W}_{eq} normalized by N , the number of uses of the underlying channel \mathbf{V} . Here, $C(\mathbf{Q}) = \max_{\mathbf{p}} I(\mathbf{p}, \mathbf{Q})$ where $I(\mathbf{p}, \mathbf{Q})$ is the mutual information between the input \mathbf{X} and the output \mathbf{Y} when $\mathbf{X} \sim \mathbf{p}$ and $\mathbf{Y}|\mathbf{X} \sim \mathbf{Q}$.

In this paper, we will make connections between $C_{N,L}(\mathbf{V})$ and the zero error capacity of the underlying channel \mathbf{V} . Recall that the zero error capacity is defined as the maximum rate at which communication is possible with zero probability of error. The notion of zero error capacity was introduced in [7] (see [16] for further details). It can be computed as follows. For a channel with transition matrix \mathbf{V} , we call two input letters k and ℓ adjacent if there exists an output letter j such that $[\mathbf{V}]_{k,j} > 0$ and $[\mathbf{V}]_{\ell,j} > 0$. We then construct a graph $G(\mathbf{V})$ corresponding to the stochastic matrix \mathbf{V} having as vertex set the possible inputs of \mathbf{V} and in which two edges are connected if the corresponding input letters are adjacent. Denote by $M_0(\mathbf{V})$ the largest number of vertices in $G(\mathbf{V})$ no two of which are connected by an edge (or, equivalently, the largest number of input letters of \mathbf{V} no two of which are adjacent). In graph theory, $M_0(\mathbf{V})$ is called the *independence number* of $G(\mathbf{V})$. The zero error capacity of \mathbf{V} is then given by

$$C_0(\mathbf{V}) \triangleq \sup_n \frac{1}{n} \log M_0(\mathbf{V}^{\otimes n}). \quad (2)$$

Clearly, for any DMC with transition probability matrix \mathbf{V} , any intermediate processing of length N and any network length L , we have

$$\frac{1}{N} M_0(\mathbf{V}^{\otimes N}) \leq C_{N,L}(\mathbf{V}) \leq C(\mathbf{V}), \quad (3)$$

where the lower bound is achievable by using the same length- N zero error code at each node in the network and the upper bound is the network min-cut capacity [1].

B. Main Results

Our main result states that the capacity of an infinite cascade of identical DMCs with transition probability matrix \mathbf{V} , and with intermediate nodes restricted to process only blocks of N symbols, cannot exceed the zero error capacity of the underlying channel. More precisely, it states that the lower bound in (3) is tight as $L \rightarrow \infty$.

Theorem II.1. *There exists an optimal processing $\mathbf{M}^* \in \mathcal{S}_{|\mathcal{Y}|^N, |\mathcal{X}|^N}$ performed at all the relays such that*

$$\begin{aligned} \lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) &= \lim_{L \rightarrow \infty} \frac{1}{N} C\left(\left(\mathbf{M}^* \mathbf{V}^{\otimes N}\right)^L\right) \\ &= \frac{1}{N} \log M_0(\mathbf{V}^{\otimes N}) \leq C_0(\mathbf{V}). \end{aligned}$$

This theorem summarizes a number of results presented in Section V. It tells us that the limit $\lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V})$ exists and can be achieved using identical processing at the relays. Moreover, the optimal identical processing \mathbf{M}^* in the limit for large L corresponds to the best, in the sense of highest rate, zero error code of length N for the channel \mathbf{V} . Hence, the capacity of the infinite cascade equals the rate of this zero error code and is always upper bounded by the zero error capacity of \mathbf{V} . Notice that any rate strictly below the zero-error capacity is achievable with finite length processing [16]. Thus this result has a nice intuitive interpretation: as $L \rightarrow \infty$, the zero error capacity is the only part of the transmitted information rate that we may hope to preserve.

The next result illustrates the behavior of long cascades of identical channels, arising for example when the intermediate nodes perform the same processing.

Theorem II.2. *For any square stochastic matrix \mathbf{Q}*

$$\lim_{L \rightarrow \infty} C(\mathbf{Q}^L) = \log D(\mathbf{Q}),$$

where $D(\mathbf{Q})$ is the number of eigenvalues of modulus (magnitude) one of \mathbf{Q} . Furthermore, define $\widehat{\mathbf{Q}}$ as the stochastic matrix obtained by deleting from \mathbf{Q} all the inessential indices¹ and $\lambda_2(\mathbf{Q})$ as the second largest eigenvalue modulus of \mathbf{Q} , then the limiting capacity is achieved exponentially fast, with exponent bounded as

$$-\log |\lambda_2(\mathbf{Q})| \leq \liminf_{L \rightarrow \infty} -\frac{1}{L} \log \left(C(\mathbf{Q}^L) - \log D(\mathbf{Q}) \right) \leq -2 \log |\lambda_2(\widehat{\mathbf{Q}})|$$

Both bounds for the exponent are tight. In particular, the upper bound is tight if $\mathbf{Q} = \widehat{\mathbf{Q}}$.

This theorem summarizes a number of results presented in Section IV. It tells us that the limit $\lim_{L \rightarrow \infty} C(\mathbf{Q}^L)$ exists and can be easily computed as the logarithm on the number of eigenvalues

¹For an exposition on inessential indices, see our short review in Section IV-A.

of modulus one of \mathbf{Q} . Furthermore, convergence to the limiting capacity is exponentially fast in L , at a rate that depends on the second largest eigenvalue modulus of \mathbf{Q} . This implies that even for long, but not infinite, cascades the derived limiting results are meaningful.

For finite L and $N \rightarrow \infty$ the relays can use a capacity achieving code and communicate reliably as long as the rate of this code is below the capacity of the channel \mathbf{V} . That no other coding strategy can do better than this is clear from either the min-cut bound [1] or directly from the data processing inequality. Hence the capacity of the cascaded channel with infinite complexity processing at the relays is

$$C_{\infty,L}(\mathbf{V}) \triangleq \lim_{N \rightarrow \infty} C_{N,L}(\mathbf{V}) = C(\mathbf{V}),$$

i.e., the upper bound in (3) is tight. From Theorem II.1, we have that for $L \rightarrow \infty$ and finite N

$$C_{N,\infty}(\mathbf{V}) \triangleq \lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) \leq C_0(\mathbf{V}),$$

i.e., the lower bound in (3) is tight. The limits $C_{\infty,L}(\mathbf{V})$ and $C_{N,\infty}(\mathbf{V})$ might differ quite substantially. A natural question to ask is, what happens if both N and L are allowed to grow. Our last result tells us how fast N needs to grow with L in order to achieve rates above the zero error capacity.

Theorem II.3. *Consider a cascade of L identical channels \mathbf{V} where intermediate nodes can perform processing of length N . A processing of length²*

$$N = \Theta(\log(L))$$

is sufficient to achieve $C_{N,L}(\mathbf{V}) \geq (1-\alpha)C_0(\mathbf{V}) + \alpha C(\mathbf{V})$ for all $\alpha \in [0, 1]$. Furthermore, there exists $\beta \geq 0$ such that logarithmic growth is necessary for any $\alpha \geq \beta$.

This theorem summarizes a number of results presented in Section VII. The derivation of these results is founded on upper and lower bounds on $C_{N,L}(\mathbf{V})$ (derived in Section VI) that are valid for all values of N and L . These bounds have merit of their own, as they increase our understanding of the expected achievable rates for finite values of N and L .

²We use Knuth's notation: $f(n) = O(g(n))$ means that there exists a constant c and integer n_0 such that $f(n) \leq cg(n)$ for $n > n_0$; $f(n) = \Theta(g(n))$ denotes that $f(n) = O(g(n))$ as well as $g(n) = O(f(n))$.

From Theorem II.1 and [16] we know that, for any network length L , any rate below the zero error capacity can be achieved with a processing of finite length N . Theorem II.3 on the other hand tells us that, for rates above the zero error capacity, N needs to increase at most logarithmically with the length of the network L . Moreover, in many cases of interest, logarithmic growth is also necessary. This is the case in the following example that illustrates the use of Theorems II.1, II.2, and II.3.

Example II.1. Line Network of Binary Symmetric Channels

Consider a cascade of L BSC(p) (Binary Symmetric Channels with crossover probability p), where, without loss of generality, $p \in [0, 1/2]$. The capacity of a BSC(p) is

$$C(\mathbf{V}) = \log(2) + p \log p + (1 - p) \log(1 - p) \triangleq C(p),$$

and the zero error capacity is $C_0(\mathbf{V}) = 0$.

From Theorems II.1 and II.2, we have that $C_{N,\infty}(\mathbf{V}) = 0$, that is, as L becomes very large finite length processing does not offer any benefits. In other words, whether we use finite length processing or no processing at all, we cannot convey any information rate from the source to the destination. Moreover, if all intermediate nodes perform the same processing then this limit is reached exponentially fast in L .

Now assume that we need to convey some information through this network. The next question is how N should scale with L in order to achieve a strictly positive rate. Theorem II.3 tells us that $N = \Theta(\log L)$ is *sufficient* to achieve any positive fraction of the min-cut capacity. Furthermore, as we will show in Example VII.2, for a network of BSCs the constant β in Theorem II.3 is equal to zero, and hence $N = \Theta(\log L)$ is also *necessary* to achieve *any* positive fraction of the min-cut capacity.

Note that, for finite L , finite complexity processing at the intermediate nodes can benefit the overall end-to-end achievable rate. To see this, consider the case where the intermediate nodes simply *forward* the incoming bits to the next node without further processing. For a cascade of BSC(p), forwarding is clearly the optimal processing for $N = 1$. In fact, since a cascade of L BSC(p) is itself a BSC with parameter $p_L \triangleq \frac{1 - (1 - 2p)^L}{2}$ [1], the end-to-end achievable rate for $N = 1$ is

$$C_{1,L}(\mathbf{V}) = C(\mathbf{V}^L) = C\left(\frac{1 - (1 - 2p)^L}{2}\right)$$

Forwarding is also the optimal processing for $N = 2$, that is $C_{2,L}(\mathbf{V}) = C_{1,L}(\mathbf{V})$. Figure 2 illustrates, for a cascade of $L = 2$ BSC(p), the forwarding capacity $C_{1,2}(\mathbf{V})$ and the min-cut capacity $C(\mathbf{V})$.

The same figure also shows the achievable end-to-end rate when the intermediate node decodes and re-encodes a repetition code of length $N = 3$. By exhaustive search, for $L = 2$ and $N = 3$, the optimal processing can be found to be either forwarding (small p) or repetition coding (large p). In particular, repetition coding achieves a rate 1.7 times higher than forwarding for $p \rightarrow 1/2$.

In Example III.1 we will analytically calculate the benefits that repetition coding can offer. We will see that this very simple processing allows to bridge a significant percentage of the gap between forwarding (no processing) and optimal processing.

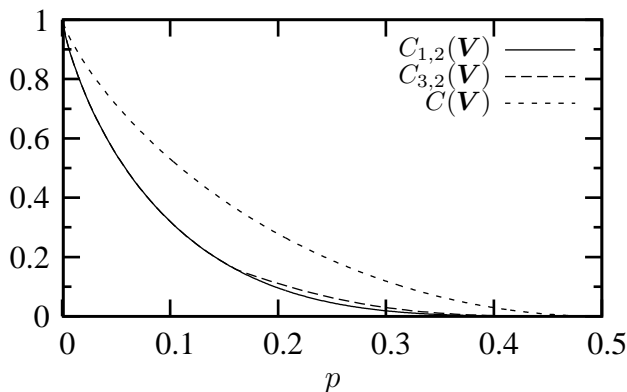


Fig. 2. Capacity of $L = 2$ cascaded BSC(p) with simple forwarding $C_{1,2}(\mathbf{V})$, optimal intermediate processing of blocklength three $C_{3,2}(\mathbf{V})$, and optimal infinite length processing at the relay $C(\mathbf{V})$.

◇

C. Detailed Outline

In Section III, we derive properties that the optimal processing must satisfy for all values of N and L . Namely, we show that the optimal processings are deterministic functions that can be thought of as a decoding and a re-encoding operation.

We then derive results for finite N and $L \rightarrow \infty$. A key result we use is that, in this case, identical processing is optimal. As a consequence, we start in Section IV by deriving the capacity of an infinite cascade of identical channels *with identical processing* at the relay nodes. We

show that the limiting capacity can be easily calculated in terms of the number of eigenvalues of modulus one of the channel transition probability matrix of the underlying DMC. We also show that convergence to this limiting value is exponential in the length of the cascade and find tight bounds on the exponent in terms of the second largest eigenvalue modulus of the channel transition probability matrix.

Then, in Section V, we consider cascades *with finite length processing* at the relays. We prove that without loss of optimality the optimal intermediate processing at the relays can be chosen to be identical. By using the results in Section IV, we show that the limiting capacity of the cascade equals the rate of the best zero error code with blocklength N for the channel the cascade is composed of, thus giving a complete characterization of infinitely long networks.

In Section VI, we derive upper and lower bounds on the capacity of the cascade with optimal intermediate processing valid for any L and N . We then build on those results to analyze the case where both $N \rightarrow \infty$ and $L \rightarrow \infty$. In Section VII we calculate how fast the blocklength N of the intermediate processing should increase with the network length L , to achieve rates strictly above the zero error capacity. We use results from Section VI to show that logarithmic scaling of N with L is sufficient, and in many cases of interest also necessary. Finally, we conclude the paper in Section VIII.

III. PROPERTIES OF OPTIMAL PROCESSING

We will call a matrix *binary* if its components take values in $\{0, 1\}$. For a binary and stochastic matrix each row contains exactly one 1 and all other components are 0. Every binary stochastic matrix \mathbf{M} of dimension $|\mathcal{Y}|^N \times |\mathcal{X}|^N$ defines a (deterministic) mapping $f(\cdot)$ from \mathcal{Y}^N to \mathcal{X}^N , where $f(\mathbf{y}) = \mathbf{x}$ if and only if $[\mathbf{M}]_{\mathbf{x}, \mathbf{y}} = 1$. We will show that the stochastic matrices $\{\mathbf{M}_i^*\}_{i=1}^{L-1}$, describing the optimal processing at the relay nodes, can be chosen to be binary and stochastic. Hence the relay nodes should employ a deterministic mapping.

The following two results are valid for arbitrary line networks, not only cascades of the same DMC.

Proposition III.1. *There exists an optimal processing $\{\mathbf{M}_i^*\}_{i=1}^{L-1}$ defining deterministic mappings at each relay, i.e., such that every \mathbf{M}_i^* is binary and stochastic.*

Proof: To simplify notation, we will consider one relay only and drop the corresponding

subscripts of M . It is straightforward to extend the proof for the general case of L channels.

For any fixed input distribution \mathbf{p} , the mutual information $I(\mathbf{p}, \mathbf{W}_{\text{eq}}(M))$ is a convex function of the channel matrix $\mathbf{W}_{\text{eq}}(M)$ as defined in (1) [17]. As $\mathbf{W}_{\text{eq}}(M)$ is a linear function of M , $I(\mathbf{p}, \mathbf{W}_{\text{eq}}(M))$ is also convex in M . Since the set $\mathcal{M} \triangleq \mathcal{S}_{|\mathcal{Y}|^N, |\mathcal{X}|^N}$ of all possible intermediate processing is convex, bounded and closed, the maximum of the convex function $I(\mathbf{p}, \mathbf{W}_{\text{eq}}(M))$ is attained at an extreme point of \mathcal{M} [18]. Consider a matrix $M \in \mathcal{M}$ which is not binary. Then, there must exist two tuples (i, j) and (i, k) such that $[M]_{i,j}, [M]_{i,k} \in (0, 1)$. Define B to be the matrix with all zero entries except $[B]_{i,j} = 1$ and $[B]_{i,k} = -1$, and let

$$M^+ \triangleq M + \epsilon B$$

$$M^- \triangleq M - \epsilon B$$

for some $\epsilon > 0$ such that $M^+, M^- \in \mathcal{M}$. Then, since the matrix M can be written as a linear combination of M^+ and M^- , it cannot be an extreme point of \mathcal{M} . Hence, the only extreme points of \mathcal{M} are binary stochastic matrices. ■

Proposition III.1 allows us to restrict our attention to deterministic mappings from $|\mathcal{Y}|^N$ to $|\mathcal{X}|^N$ at the relay nodes without loss of optimality. The next proposition shows that we can view every binary stochastic matrix M of rank ρ to be composed of a decoder followed by an encoder, that is, $M = M_D M_E$, where both M_D and M_E are binary stochastic matrices of dimension $|\mathcal{Y}|^N \times \rho$ and $\rho \times |\mathcal{X}|^N$, respectively. The matrix M_D can be interpreted as a decoder, mapping a length- N sequence at the output of one channel into one out of ρ possible “messages”. The matrix M_E can be seen as an encoder, mapping these “messages” back into a length- N sequence to be used as input for the next channel. We will give a simple constructive proof to show that this decomposition is always possible.

Proposition III.2. *For every binary stochastic matrix M of dimension $|\mathcal{Y}|^N \times |\mathcal{X}|^N$ and of rank ρ , there exists a pair of binary stochastic matrices M_D and M_E of dimension $|\mathcal{Y}|^N \times \rho$ and $\rho \times |\mathcal{X}|^N$, respectively, such that $M = M_D M_E$.*

Proof: As the stochastic matrix M is binary and has rank ρ , there are exactly ρ nonzero columns. Let $\{i_1, \dots, i_\rho\}$ denote the positions at which these nonzero columns occur. Call \mathbf{m}_i the i -th column of M and \mathbf{e}_i the i -th unit vector of length $|\mathcal{X}|^N$ (i.e., the vector containing all

zeros except a 1 at position i). The desired decomposition is then

$$\mathbf{M} = \begin{pmatrix} \mathbf{m}_{i_1} & \cdots & \mathbf{m}_{i_\rho} \end{pmatrix} \begin{pmatrix} \mathbf{e}_{i_1}^T \\ \vdots \\ \mathbf{e}_{i_\rho}^T \end{pmatrix} \triangleq \mathbf{M}_D \mathbf{M}_E.$$

■

The following example shows a decoding and encoding scheme for the relay nodes of a line network of $\text{BSC}(p)$. It shows how finite complexity intermediate node processing can benefit the overall end-to-end achievable rate.

Example III.1. Repetition Coding for Line Networks of Binary Symmetric Channels

Consider a cascade of L $\text{BSC}(p)$, as in Example II.1. Assume that the processing length N is an odd integer and that the relays send either the all-one sequence if the received word has Hamming weight larger than $N/2$, or the all-zero sequence otherwise. In other words, the processing at the intermediate nodes corresponds to the decoding and re-encoding of a *repetition code* of length N . The encoding and decoding matrices have the form

$$\mathbf{M}_E = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \in \mathcal{S}_{2,2^N}, \text{ and } \mathbf{M}_D = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix} \in \mathcal{S}_{2^N,2},$$

where the inputs and outputs of the channels are ordered in terms of increasing Hamming weight, and $\mathbf{0}$ and $\mathbf{1}$ in \mathbf{M}_D are, respectively, the all-zero and all-ones column vectors of dimension 2^{N-1} . If we assume that the destination node also uses the decoder of the repetition code, then this communication strategy turns the original cascade of L $\text{BSC}(p)$ into a cascade of L $\text{BSC}(p_N)$, where p_N is the probability that more than $N/2$ bit flips have occurred in the transmission of N bits, i.e.,

$$p_N \triangleq \sum_{j=(N+1)/2}^N \binom{N}{j} p^j (1-p)^{N-j}.$$

Notice that for odd N

$$1 - 2p_N = \sum_{j=0}^{(N-1)/2} \beta_{j,N} (1-2p)^{2j+1}$$

for some values $\beta_{0,N}, \beta_{1,N}, \dots, \beta_{(N-1)/2,N}$, and where

$$\beta_{0,N} = \lim_{p \rightarrow 1/2} \frac{1 - 2p_N}{1 - 2p} = \frac{N}{2^{N-1}} \binom{N-1}{(N-1)/2} \approx \sqrt{2N/\pi} \text{ for } N \gg 1.$$

Hence, the gain of the length- N repetition coding scheme over forwarding for a line network of L BSC(p) with large cross-over probability p is

$$\lim_{p \rightarrow 1/2} \frac{\frac{1}{N} C\left(\frac{1-(1-2p)^L}{2}\right)}{C\left(\frac{1-(1-2p)^L}{2}\right)} = \frac{(\beta_{0,N})^{2L}}{N} \approx \left(\frac{2}{\pi}\right)^L N^{L-1} \quad \text{for } N \gg 1,$$

since $C(p) \approx (1-2p)^2/2$ for $p \rightarrow 1/2$.

◇

IV. CAPACITY OF A CASCADE OF IDENTICAL CHANNELS

In the next section, we will prove that *identical* processing is optimal for an infinitely long cascade of channels. In this section, we hence derive the capacity of an infinite cascade of a DMCs with input and output alphabets of the same cardinality and transition matrix \mathbf{Q} , that is, we compute

$$\lim_{L \rightarrow \infty} C(\mathbf{Q}^L). \quad (4)$$

In other words, the channel \mathbf{Q} is cascaded with itself *without any intermediate processing*, or \mathbf{Q} can be the result of applying the *same intermediate processing* at each relay, that is $\mathbf{Q} = \mathbf{M}_E \mathbf{W} \mathbf{M}_D$.

In Section IV-A we briefly review the canonical decomposition of non-negative matrices, and we compute the limit of \mathbf{Q}^L as $L \rightarrow \infty$ and the rate of convergence of \mathbf{Q}^L to its limiting value for the class of stochastic matrices of interest to compute (4). We use these results in Section IV-B to characterize the limiting capacity of an arbitrary channel cascaded L times with itself, and in Section IV-C to characterize the rate of convergence to the limiting value.

A. Canonical Form of Stochastic Matrices

Our exposition closely follows [19]. Let \mathbf{Q} be a square stochastic matrix and denote by $\mathcal{J} \triangleq \{1, \dots, m\}$ the set of its (row and column) indices. We say that the index i *leads* to index j , and write $i \rightarrow j$, if $[\mathbf{Q}^k]_{i,j} > 0$ for some $k \geq 1$. If $i \rightarrow j$ and $j \rightarrow i$, we say that i and j *communicate*. An index i is called *essential* if $i \rightarrow j$ implies $j \rightarrow i$. If i is not essential, it is called *inessential*. This partitions the set of indices \mathcal{J} into the set of essential indices \mathcal{E} and inessential indices \mathcal{I} . The set of essential indices \mathcal{E} can furthermore be partitioned into communicating classes \mathcal{C} , such that all indices communicating with each other are in the same

class. The canonical form of a matrix \mathbf{Q} is obtained by relabeling its indices in such a way that all indices of the same essential communicating class are consecutive, and every inessential index is greater than any essential index. Formally, this corresponds to pre- and post-multiplying \mathbf{Q} by some permutation matrix $\mathbf{\Pi}$. This results in a matrix of the canonical form

$$\tilde{\mathbf{Q}} = \mathbf{\Pi}\mathbf{Q}\mathbf{\Pi}^T = \begin{pmatrix} \mathbf{P}_1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_{|\mathcal{C}|} & \mathbf{0} \\ \mathbf{R}_1 & \mathbf{R}_2 & \cdots & \mathbf{R}_{|\mathcal{C}|} & \mathbf{S} \end{pmatrix}. \quad (5)$$

The square matrix \mathbf{P}_i in (5) contains the transition probabilities within the i -th essential communicating class, the square matrix \mathbf{S} the transition probabilities between the inessential indices \mathcal{I} , and the (not necessarily square) matrices \mathbf{R}_i the transition probabilities from the inessential indices to the i -th essential communicating class. The submatrices $\{\mathbf{P}_i\}_{i=1}^{|\mathcal{C}|}$ are by construction *irreducible* [19].

The *period* of an index i is defined as the greatest common divisor of those k for which $[\mathbf{Q}^k]_{i,i} > 0$. All indices in the same communicating class have the same period, which is referred to as the period of the class. Denote by d_i the period of the submatrix \mathbf{P}_i . If $d_i = 1$, then \mathbf{P}_i is called *primitive*, i.e., it is irreducible and aperiodic. If $d_i > 1$, then \mathbf{P}_i can be written in a canonical form (again by permuting indices) such that, for any integer ℓ , $\mathbf{P}_i^{d_i \ell}$ is of the form

$$\mathbf{P}_i^{d_i \ell} = \begin{pmatrix} \mathbf{P}_{i,1}^\ell & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{i,2}^\ell & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_{i,d_i}^\ell \end{pmatrix}, \quad (6)$$

where the square matrices $\{\mathbf{P}_{i,j}\}_{j=1}^{d_i}$ on the main diagonal are *primitive*.

The following lemma gives the limiting expression of \mathbf{Q}^L when $L \rightarrow \infty$ for certain \mathbf{Q} . As we shall see, the class of \mathbf{Q} covered by the theorem is general enough for the purposes of computing the capacity in (4).

Lemma IV.1. *Let \mathbf{Q} be a square stochastic matrix with $|\mathcal{C}|$ aperiodic essential communicating classes, i.e., \mathbf{Q} is in the canonical form given in (5) with all its diagonal irreducible submatrices*

$\{\mathbf{P}_i\}_{i=1}^{|\mathcal{C}|}$ with period one, then

$$\mathbf{Q}^\infty \triangleq \lim_{L \rightarrow \infty} \mathbf{Q}^L = \begin{pmatrix} \mathbf{1}\pi_1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}\pi_2 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}\pi_{|\mathcal{C}|} & \mathbf{0} \\ \mathbf{a}_1\pi_1 & \mathbf{a}_2\pi_2 & \cdots & \mathbf{a}_{|\mathcal{C}|}\pi_{|\mathcal{C}|} & \mathbf{0} \end{pmatrix}, \quad (7)$$

where the row vector π_i is the unique stationary distribution of \mathbf{P}_i , and column vector $\mathbf{a}_i \triangleq (\mathbf{I} - \mathbf{S})^{-1} \mathbf{R}_i \mathbf{1}$ (\mathbf{I} indicates the identity matrix and $\mathbf{1}$ the column vector of all ones).

Proof: The proof follows from [19, Th. 4.1, Th. 4.2, Th. 4.3, Th. 4.7]. ■

The speed of convergence to the limiting expression in (7) is exponential in L , with the exponent related to $\{\lambda_i(\mathbf{Q})\}_{i=1}^s$, the ordered (written without repetition) eigenvalues of \mathbf{Q} such that

$$|\lambda_1(\mathbf{Q})| > |\lambda_2(\mathbf{Q})| > \dots > |\lambda_s(\mathbf{Q})|.$$

The next lemma provides detailed information about the speed at which convergence in Lemma IV.1 takes place.

Lemma IV.2. *Let \mathbf{Q} be a square stochastic matrix with only aperiodic essential communicating classes. The entries of \mathbf{Q}^k converge to the entries of \mathbf{Q}^∞ in (7) exponentially fast³ in k , with exponent not smaller than $-\log |\lambda_2(\mathbf{Q})|$ and with equality for at least one entry of \mathbf{Q} .*

Proof: From [19, Th. 1.2, Th. 4.2], the entries of \mathbf{P}_i^k converge to the entries of $\mathbf{1}\pi_i$ exponentially fast, all with exponent $-\log |\lambda_2(\mathbf{P}_i)|$.

From [19, Th. 4.3] and [20, Cor. 5.6.14], the entries of \mathbf{S}^k converge to zero exponentially fast, with exponent not smaller than $-\log |\lambda_1(\mathbf{S})|$, and with equality for at least one entry of \mathbf{S} . In other words, $|\lambda_1(\mathbf{S})|$ is the spectral radius of \mathbf{S} and $|\lambda_1(\mathbf{S})| < 1$.

From [19, Th. 4.7], the entries of $\mathbf{R}_i^{(k)} = \sum_{j=1}^{k-1} \mathbf{S}^j \mathbf{R}_i \mathbf{P}_i^{k-1-j}$ (the position of the matrix $\mathbf{R}_i^{(k)}$ in \mathbf{Q}^k is the same as the position of the matrix \mathbf{R}_i in \mathbf{Q}) converge to the entries of $\mathbf{a}_i\pi_i$ exponentially fast, with exponent not smaller than $-\log \max\{|\lambda_2(\mathbf{P}_i)|, |\lambda_1(\mathbf{S})|\}$.

³ $f(k)$ converges exponentially to a with exponent b if $\lim_{k \rightarrow \infty} f(k) = a$ and $\liminf_{k \rightarrow \infty} -\frac{1}{k} \log (f(k) - a) = b$.

Hence, all entries of \mathbf{Q} converge to their respective limit with exponent not smaller than

$$-\log \max\{|\lambda_1(\mathbf{S})|, |\lambda_2(\mathbf{P}_1)|, \dots, |\lambda_2(\mathbf{P}_{|\mathcal{C}|})|\}, \quad (8)$$

and with equality for at least one entry of \mathbf{Q} . Since the eigenvalues of the block lower triangular matrix \mathbf{Q} are the union of the eigenvalues of its diagonal blocks [21, Ex. 4, page 64], and since $\lambda_1(\mathbf{Q}) = 1$ because \mathbf{Q} is a stochastic matrix, it follows that (8) equals $-\log |\lambda_2(\mathbf{Q})|$, thus yielding the desired result. ■

An immediate consequence of Lemma IV.1 is that:

Corollary IV.3. *Let \mathbf{Q} be a square stochastic matrix with $|\mathcal{C}|$ aperiodic essential communicating classes. Then*

$$|\mathcal{C}| = \text{rank}\left(\lim_{L \rightarrow \infty} \mathbf{Q}^L\right) = D(\mathbf{Q}), \quad (9)$$

where $D(\mathbf{Q})$ is the multiplicity (algebraic and also geometric) of $\lambda_1(\mathbf{Q}) = 1$.

Proof: That the rank of $\lim_{L \rightarrow \infty} \mathbf{Q}^L$ is $|\mathcal{C}|$ follows immediately from inspection of (7).

That $|\mathcal{C}|$ is also the multiplicity of $\lambda_1(\mathbf{Q}) = 1$ is a consequence of the following facts. First, the eigenvalues of a block lower triangular matrix are the union of the eigenvalues of its block diagonal blocks [21, Ex. 4, page 64]. Second, each primitive stochastic matrix \mathbf{P}_i on the main diagonal of \mathbf{Q} has only one eigenvalue of maximum modulus ($\lambda_1(\mathbf{P}_i) = 1$). This eigenvalue, referred to as *Perron-Frobenius root*, has algebraic and geometric multiplicity one [19, Th. 1.1]. Third, all the eigenvalues of \mathbf{S} are in modulus strictly less than 1. Hence, the eigenvalue of \mathbf{Q} of maximum modulus ($\lambda_1(\mathbf{Q}) = 1$) has multiplicity $D(\mathbf{Q})$ given by the number of primitive stochastic matrices \mathbf{P}_i on the main diagonal of \mathbf{Q} , which is $|\mathcal{C}|$ by definition.

Notice that equation (9) implies that the eigenvalues of $\lim_{L \rightarrow \infty} \mathbf{Q}^L$ are either zero or one. ■

The following example illustrates these definitions.

Example IV.1. Let $p \in (0, 1)$, $t \in (0, 1)$, and consider

$$\mathbf{Q} = \begin{pmatrix} 1-p & 0 & 0 & p \\ 0 & 1 & 0 & 0 \\ 0 & t & 1-t & 0 \\ p & 0 & 0 & 1-p \end{pmatrix},$$

The matrix \mathbf{Q} has essential indices $\mathcal{E} = \{1, 2, 4\}$, inessential indices $\mathcal{I} = \{3\}$, and two essential communicating classes $\mathcal{C} = \{\{1, 4\}, \{2\}\}$. The canonical form $\tilde{\mathbf{Q}}$ is

$$\tilde{\mathbf{Q}} = \begin{pmatrix} 1-p & p & 0 & 0 \\ p & 1-p & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & t & 1-t \end{pmatrix} \quad (10)$$

with

$$\mathbf{P}_1 = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}, \quad \mathbf{P}_2 = \begin{pmatrix} 1 \end{pmatrix},$$

and

$$\mathbf{R}_1 = \begin{pmatrix} 0 & 0 \end{pmatrix}, \quad \mathbf{R}_2 = \begin{pmatrix} t \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 1-t \end{pmatrix}.$$

Both matrices \mathbf{P}_1 and \mathbf{P}_2 have period 1. Hence the matrix \mathbf{Q} falls into the category covered by Lemma IV.1, Lemma IV.2 and Corollary IV.3. The eigenvalues of \mathbf{Q} are $\{1, 1-2p, 1-t\}$ and $\lambda_1(\mathbf{Q}) = 1$ has multiplicity $D(\mathbf{Q}) = 2$.

From Lemma IV.1

$$\tilde{\mathbf{Q}}^\infty = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad (11)$$

and, from Lemma IV.2, the speed of convergence is exponential with exponent

$$-\log \max\{|1-2p|, 1-t\} \quad (12)$$

From Corollary IV.3, the rank of $\tilde{\mathbf{Q}}^\infty$ is $D(\mathbf{Q}) = 2$.

In this specific example, we could also have computed directly the limiting value in (11) and the exponent in (12) from the following explicit expression for \mathbf{Q}^L

$$\tilde{\mathbf{Q}}^L - \tilde{\mathbf{Q}}^\infty = \begin{pmatrix} +x/2 & -x/2 & 0 & 0 \\ -x/2 & +x/2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -y & +y \end{pmatrix}, \quad (13)$$

with $x = (1-2p)^L$ and $y = (1-t)^L$. ◇

B. Capacity of an Infinite Cascade of Identical Channels

We will now use the results from Section IV-A to find the capacity of an infinite cascade of an arbitrary channel \mathbf{Q} *without intermediate processing*.

Theorem IV.4. *Consider a square stochastic matrix \mathbf{Q} . Let $D(\mathbf{Q})$ be the number of eigenvalues on modulus 1 of \mathbf{Q} . Then*

$$\lim_{L \rightarrow \infty} C(\mathbf{Q}^L) = \log D(\mathbf{Q}). \quad (14)$$

Proof: For any square stochastic matrix \mathbf{Q} the limit $\lim_{L \rightarrow \infty} C(\mathbf{Q}^L)$ exists since the sequence $\{C(\mathbf{Q}^\ell)\}_{\ell \in \mathbb{N}}$ is non-increasing and bounded below. Indeed, by the data processing inequality and the non-negativity of mutual information, we have that $C(\mathbf{Q}) \geq C(\mathbf{Q}^2) \geq \dots \geq 0$. The existence of the limit implies that for any $d \in \mathbb{N}$

$$\lim_{L \rightarrow \infty} C(\mathbf{Q}^L) = \lim_{L \rightarrow \infty} C(\mathbf{Q}^{dL}). \quad (15)$$

By the continuity of mutual information in the channel transition probability matrix [1] and assuming that $\lim_{L \rightarrow \infty} \mathbf{Q}^{dL}$ exists (which will be shown to hold in the next paragraph)

$$\lim_{L \rightarrow \infty} C(\mathbf{Q}^{dL}) = C\left(\lim_{L \rightarrow \infty} \mathbf{Q}^{dL}\right). \quad (16)$$

Without loss of generality, assume that \mathbf{Q} is in canonical form in (5). The notation is the same as in the Section IV-A. Let d_i be the period of the irreducible matrix \mathbf{P}_i and denote by d the least common multiple of the $\{d_i\}$. By (15), we can limit our attention to the powers of \mathbf{Q}^d . From (6), we know that \mathbf{P}_i^d is block diagonal, since d is a multiple of d_i , and it has exactly d_i primitive square matrices on the main diagonal. Hence \mathbf{Q}^d has the form of (5) with $\sum_{k \geq 1} k|\mathcal{C}_k|$ aperiodic essential communicating classes, where we have denoted by \mathcal{C}_k the set of essential communicating classes with period k of \mathbf{Q} . Hence $\lim_{L \rightarrow \infty} \mathbf{Q}^{dL}$ is given by (7) in Lemma IV.1, where the number of aperiodic essential communicating classes is $\sum_{k \geq 1} k|\mathcal{C}_k|$.

Since capacity is upper bounded by the logarithm of the rank of the channel transition probability matrix [22] and from Corollary IV.3, we have that

$$C\left(\lim_{L \rightarrow \infty} \mathbf{Q}^{dL}\right) \leq \log \text{rank}\left(\lim_{L \rightarrow \infty} \mathbf{Q}^{dL}\right) = \log \sum_{k \geq 1} k|\mathcal{C}_k| = \log D(\mathbf{Q}^d) = \log D(\mathbf{Q}). \quad (17)$$

where the last equality in (17) follows since the number of eigenvalues of modulus 1 of \mathbf{Q} is the same as the number of eigenvalues of modulus one of \mathbf{Q}^k for any integer k . It is easy to

see that inequality in (17) is achievable by using one input per essential communicating class of \mathbf{Q}^d with uniform probability. This shows that (14) holds, thus concluding the proof. ■

C. Convergence to Asymptotic Value

We showed in the last section that the capacity of an infinite cascade of identical DMCs can be computed very easily in terms of the number of eigenvalues of modulus one of the channel transition probability matrix. In this section, we will see that convergence to this limiting expression is exponential in the length of the cascade. This implies that for long, but still finite, cascades of identical channels, the limiting result derived in the previous section is meaningful.

We define the exponential asymptotic rate at which capacity decays when $L \rightarrow \infty$ as

$$E_L(\mathbf{Q}) \triangleq \liminf_{L \rightarrow \infty} -\frac{1}{L} \log (C(\mathbf{Q}^L) - \log D(\mathbf{Q})).$$

The following theorem bounds $E_L(\mathbf{Q})$ in terms of the eigenvalues of the channel matrix \mathbf{Q} .

Theorem IV.5. *Let \mathbf{Q} be a stochastic matrix, and define $\widehat{\mathbf{Q}}$ as the stochastic matrix obtained by deleting from \mathbf{Q} all inessential indices. Then*

$$-\log |\lambda_2(\mathbf{Q})| \leq E_L(\mathbf{Q}) \leq -2 \log |\lambda_2(\widehat{\mathbf{Q}})|, \quad (18)$$

where $\lambda_2(\mathbf{Q})$ denotes the second largest eigenvalue modulus of the channel matrix \mathbf{Q} . Moreover, the upper bound in (18) is tight if $\mathbf{Q} = \widehat{\mathbf{Q}}$.

Proof: The proof, due to its length and technicality, is reported in Appendix I. ■

Interestingly, the speed of convergence of $C(\mathbf{Q}^L)$ in Theorem IV.5 is not necessarily the same as the speed of convergence of \mathbf{Q}^L , which was derived in Lemma IV.2 and was found to be equal to $-\log |\lambda_2(\mathbf{Q})|$.

We illustrate the use of Theorem IV.5 by further developing Example IV.1. This example shows that both the upper and the lower bounds in Theorem IV.5 are actually tight.

Example IV.2. For the channel in Example IV.1, we have $|\lambda_2(\mathbf{Q})| = \max\{|1 - 2p|, 1 - t\}$ and $|\lambda_2(\widehat{\mathbf{Q}})| = |1 - 2p|$. The matrices $\widetilde{\mathbf{Q}}$, $\widetilde{\mathbf{Q}}^L$ and $\widetilde{\mathbf{Q}}^\infty$ are given by (10), (13), and (11), respectively. Theorem IV.5 says that the convergence of $C(\widetilde{\mathbf{Q}}^L)$ to $C(\widetilde{\mathbf{Q}}^\infty) = \log(2)$ is exponentially fast in L with exponent

$$-\log \max\{|1 - 2p|, 1 - t\} \leq E_L(\widetilde{\mathbf{Q}}) \leq -\log\{(1 - 2p)^2\}. \quad (19)$$

We will now show that with the right choice of the parameters p and t , both the upper and the lower bound in (19) can be achieved. To do that, we will directly derive the speed of convergence as follows. The capacity of $\tilde{\mathcal{Q}}^L$ in (13) can be easily computed for every L as the channel is the sum of a BSC with parameter $p_L \triangleq \frac{1-(1-2p)^L}{2}$, with capacity

$$C_{\text{BSC}}(p_L) = \log(2) - \mathcal{H}(p_L) \rightarrow \frac{(1-2p)^2}{2} \quad \text{as } L \rightarrow \infty,$$

and a Z-channel with parameter $t_L \triangleq 1 - (1-t)^L$, with capacity

$$C_Z(t_L) = \log \left(1 + \exp \left(-\frac{\mathcal{H}(t_L)}{1-t_L} \right) \right) \rightarrow 1 - t_L \quad \text{as } L \rightarrow \infty.$$

We used $\mathcal{H}(p)$ to denote the binary entropy function. Hence for large L

$$\begin{aligned} C(\tilde{\mathcal{Q}}^L) &= \log \left(\exp(C_{\text{BSC}}(p_L)) + \exp(C_Z(t_L)) \right) \\ &\approx \log \left(2 + \frac{(1-2p)^{2L}}{2} + (1-t)^L \right) \\ &\approx \log 2 + \beta \max\{(1-2p)^{2L}, (1-t)^L\} \end{aligned}$$

for some constant β independent of L . We conclude that

$$E_L(\tilde{\mathcal{Q}}) = -\log \max\{(1-2p)^2, 1-t\}.$$

Thus, in the case where $(1-2p)^2 \geq (1-t)$ the upper bound in (19) becomes tight. In the case where $(1-2p)^2 < (1-t) < |1-2p|$ neither the upper nor the lower bound is tight. And finally, in the case where $(1-2p)^2 \leq |1-2p| \leq (1-t)$ the lower bound becomes tight. \diamond

V. CAPACITY OF AN INFINITE CASCADE OF CHANNELS WITH INTERMEDIATE PROCESSING

In this section, we will characterize the optimal finite length intermediate processing for an infinite cascade and establish connections with the zero error capacity. We start by showing that, similar to the usual capacity, the zero error capacity obeys a sort of data processing inequality. This result will be used in Section V-A to show that the optimal finite length processing for an infinite cascade is a zero error code.

Proposition V.1. *Consider a cascade of L channels $\{\mathcal{Q}_i\}_{i=1}^L$. Then*

$$M_0 \left(\left(\prod_{i=1}^L \mathcal{Q}_i \right)^{\otimes n} \right) \leq M_0(\mathcal{Q}_j^{\otimes n}). \quad (20)$$

Remark. Equation (20) implies that

$$C_0\left(\prod_{i=1}^L \mathbf{Q}_i\right) \leq \min_{j \in \{1, \dots, L\}} C_0(\mathbf{Q}_j).$$

Proof: By definition $M_b(\mathbf{Q}) = D$ if and only if there exists an encoder M_E and a decoder M_D such that $M_E \mathbf{Q} M_D$ is the identity matrix of dimension D . Let (M_E, M_D) be the optimal encoder and decoder for the matrix $(\prod_{i=1}^L \mathbf{Q}_i)^{\otimes n}$. By the properties of the Kronecker product [21], we have

$$I = M_E \left(\prod_{i=1}^L \mathbf{Q}_i \right)^{\otimes n} M_D = M_E \prod_{i=1}^L (\mathbf{Q}_i^{\otimes n}) M_D = M_{E,j} \mathbf{Q}_j^{\otimes n} M_{D,j},$$

where

$$M_{E,j} \triangleq M_E \prod_{i=1}^{j-1} \mathbf{Q}_i^{\otimes n}, \quad M_{D,j} \triangleq \left(\prod_{i=j+1}^L \mathbf{Q}_i^{\otimes n} \right) M_D.$$

Hence, there exists at least one pair of zero error encoder (i.e., $M_{E,j}$) and decoder (i.e., $M_{D,j}$) for the channel $\mathbf{Q}_j^{\otimes n}$ yielding the same zero error rate as $\prod_{i=1}^L \mathbf{Q}_i^{\otimes n}$, which shows the result. ■

A. Cascade of Identical Channels

The next theorem shows that for an infinite cascade of identical DMCs, identical processing at the relays is optimal. This theorem is crucial as it allows us to optimize over only one intermediate processing instead of having to optimize over an infinite sequence of processing $\{M_i\}_{i=1}^{\infty}$.

Theorem V.2. *For a cascade of L identical DMCs, identical processing at the relays is optimal as $L \rightarrow \infty$, i.e., there exists an optimal processing M^* such that*

$$\lim_{L \rightarrow \infty} \max_{\{M_i\}_{i=1}^{L-1}} \frac{1}{N} C(\mathbf{W}_{\text{eq}}(\{M_i\})) = \lim_{L \rightarrow \infty} \frac{1}{N} C((M^* \mathbf{W})^L)$$

Proof: Let $\mathbf{W} = \mathbf{V}^{\otimes N}$ be the channel transition probability matrix of the equivalent DMC between any pair of relays and, as before, $\{M_i = M_{i,D} M_{i,E}\}_{i=1}^{L-1}$ the processing at the relays. Let $\mathbf{Q}_i \triangleq M_i \mathbf{W} \in \mathcal{S}_{\mathcal{Y}^N, \mathcal{Y}^N}$ for $i = 1, \dots, L-1$. With this

$$\mathbf{W}_{\text{eq}}(\{M_i\}) = \mathbf{W} \prod_{i=1}^{L-1} \mathbf{Q}_i.$$

An *interval chain* σ of length ℓ is defined to be a sequence of intervals $\{\sigma_i\}_{i=1}^{\ell}$, where the σ_i are each integer intervals $\{l_i, \dots, r_i\}$ and have the property that $l_i = r_{i-1} + 1$ for all

$i \in \{2, \dots, \ell\}$. Consider the product $\prod_{i=1}^{L-1} \mathbf{Q}_i$ and define $\mathbf{Q}_{\sigma_j} \triangleq \prod_{i=l_j}^{r_j} \mathbf{Q}_i$ for any integer interval $\sigma_j \subseteq \{1, \dots, L-1\}$. We will use [23, Lemma 2.4], a result originally due to Erdős and Szekeres, to show that, as $L \rightarrow \infty$, there exists an interval chain σ of arbitrary length ℓ such that all \mathbf{Q}_{σ_j} are almost identical. More precisely, for every L there exists an ℓ with the property that $\ell \rightarrow \infty$ as $L \rightarrow \infty$ such that

$$C(\mathbf{W} \prod_{i=1}^{L-1} \mathbf{Q}_i) = C(\mathbf{P} \mathbf{Q}_{\sigma_1}^{\ell} \tilde{\mathbf{P}}) + \varepsilon(L) \quad (21)$$

for some stochastic matrices \mathbf{P} and $\tilde{\mathbf{P}}$ and with $\lim_{L \rightarrow \infty} |\varepsilon(L)| = 0$. Indeed, for a fixed $k \in \mathbb{N}$ construct $\hat{\mathbf{Q}}$ from \mathbf{Q} by quantizing every component of \mathbf{Q} to the closest of the points $\{j/k\}_{j=0}^k$. The set of all possible quantized matrices (which are, in general, not stochastic) has cardinality $K \triangleq (k+1)^{n^2}$, with $n \triangleq |\mathcal{Y}|^N$. By [23, Lemma 2.4], we have that if $L > \ell^K$ then there exists an interval chain σ of length ℓ such that $\hat{\mathbf{Q}}_{\sigma_1} = \hat{\mathbf{Q}}_{\sigma_j}$ for all $j \in \{1, \dots, \ell\}$. Note that $\hat{\mathbf{Q}}_{\sigma_j}$ is defined as the quantized version of \mathbf{Q}_{σ_j} and hence $\hat{\mathbf{Q}}_{\sigma_j}$ and \mathbf{Q}_{σ_j} differ component-wise by at most $1/k$. By the above argument the product $\prod_{i=1}^{\ell} \mathbf{Q}_{\sigma_i}$ and $\mathbf{Q}_{\sigma_1}^{\ell}$ differ component-wise at most by $\frac{1}{k}(an)^{\ell}$ for some constant a independent of k . By choosing k large enough we can make this difference as small as desired. As mutual information is continuous in the channel transition probability matrix we can, for any input distribution \mathbf{p} , make the difference $|I(\mathbf{p}, \mathbf{Q}_{\sigma_1}^{\ell}) - I(\mathbf{p}, \prod_{i=1}^{\ell} \mathbf{Q}_{\sigma_i})|$ also as small as desired. Since ℓ is arbitrary, the result follows.

By the data processing inequality, we have that the capacity in (21) can be upper bounded by

$$C(\mathbf{W} \prod_{i=1}^{L-1} \mathbf{Q}_i) \leq C(\mathbf{Q}_{\sigma_1}^{\ell}) + \varepsilon(L).$$

But any stochastic matrix \mathbf{Q}_{σ_1} resulting from this procedure can be written as the product $\mathbf{M}\mathbf{W}$ for some stochastic matrix \mathbf{M} and hence $\mathbf{Q}_{\sigma_1}^{\ell}$ can be constructed from a cascade of ℓ channels \mathbf{W} by using the same processing at each relay. Hence as $L \rightarrow \infty$ (and therefore also $\ell \rightarrow \infty$) we can restrict our attention to identical processing at the relays. ■

With this last result, we are now in the position to find the optimal intermediate processing of blocklength N for an infinite cascade of identical channels and to compute the resulting capacity of the cascade. Theorem V.3 shows that the optimal intermediate processing at the relays corresponds to using the *best zero error code of blocklength N for the channel \mathbf{V}* . The resulting capacity of the cascade equals the rate of this zero error code.

Theorem V.3. *The capacity of an infinite cascade of identical DMCs with channel matrix \mathbf{V} with optimal intermediate processing of finite length N is*

$$\lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) = \frac{1}{N} \log M_0(\mathbf{V}^{\otimes N}).$$

Proof: The main steps of the proof are as follows

$$C_{N,\infty}(\mathbf{V}) = \lim_{L \rightarrow \infty} \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L) \quad (22a)$$

$$= \max_{\mathbf{M}_E, \mathbf{M}_D} \lim_{L \rightarrow \infty} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L) \quad (22b)$$

$$= \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} C\left(\lim_{L \rightarrow \infty} (\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^{dL}\right) \quad (22c)$$

$$= \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} \log D(\mathbf{M}_E \mathbf{W} \mathbf{M}_D) \quad (22d)$$

$$= \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} \log M_0\left(\lim_{L \rightarrow \infty} (\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^{dL}\right) \quad (22e)$$

$$= \frac{1}{N} \log M_0(\mathbf{W}). \quad (22f)$$

Equality in (22a) follows from Theorem V.2. The first step is to show that the limit and the maximization operation in (22a) can be exchanged, thus giving (22b). We postpone the proof of this technical step to a later stage. Equality in (22c) and in (22d) follow from Theorem IV.4, where d is the least common multiple of the periods of the essential communicating classes of $\mathbf{M}_E \mathbf{W} \mathbf{M}_D$, and where $D(\mathbf{M}_E \mathbf{W} \mathbf{M}_D)$ is the number of eigenvalues of modulus one of $\mathbf{M}_E \mathbf{W} \mathbf{M}_D$. From Lemma IV.1, the limiting channel $\lim_{L \rightarrow \infty} (\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^{dL}$ has $D(\mathbf{M}_E \mathbf{W} \mathbf{M}_D)$ non-adjacent inputs, hence (22d) is the rate of the best zero error code of length one for the limiting channel $\lim_{L \rightarrow \infty} (\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^{dL}$, thus giving (22e). Finally, equality in (22f) follows by applying Proposition V.1, which states that

$$M_0\left(\lim_{L \rightarrow \infty} (\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^{dL}\right) \leq M_0(\mathbf{W})$$

with equality if $(\mathbf{M}_E, \mathbf{M}_D)$ is an optimal zero error code for $\mathbf{W} = \mathbf{V}^{\otimes N}$, as in this case $(\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^\ell = \mathbf{I}^\ell = \mathbf{I}$ for any integer ℓ .

In order to complete the proof, we need to show that the limit and the maximization operation in (22a) can be interchanged. We have

$$\lim_{L \rightarrow \infty} \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L) \geq \lim_{L \rightarrow \infty} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L) \quad (23)$$

for any $(\mathbf{M}_E, \mathbf{M}_D)$ pair. In particular this is also true for the one maximizing the right hand side of (23) yielding

$$\lim_{L \rightarrow \infty} \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L) \geq \max_{\mathbf{M}_E, \mathbf{M}_D} \lim_{L \rightarrow \infty} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L).$$

On the other hand, the capacity of the cascade is decreasing in L and hence

$$\lim_{L \rightarrow \infty} \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L) \leq \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L).$$

By Theorem IV.4, $\lim_{L \rightarrow \infty} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L)$ exists for all choices of $(\mathbf{M}_E, \mathbf{M}_D)$ and thus for every $\varepsilon \geq 0$ there exists a $L_0 = L_0(\mathbf{M}_E, \mathbf{M}_D)$ such that

$$C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L) \leq \lim_{L \rightarrow \infty} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L) + \varepsilon \quad (24)$$

for $L \geq L_0$. By Proposition III.1, there are only finitely many $(\mathbf{M}_E, \mathbf{M}_D)$ pairs over which we optimize, hence there is a finite $\tilde{L}_0 \triangleq \max_{\mathbf{M}_E, \mathbf{M}_D} L_0(\mathbf{M}_E, \mathbf{M}_D)$ such that (24) holds for all $(\mathbf{M}_E, \mathbf{M}_D)$ pairs simultaneously if $L \geq \tilde{L}_0$. As ε is arbitrary this shows that

$$\lim_{L \rightarrow \infty} \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L) = \max_{\mathbf{M}_E, \mathbf{M}_D} \lim_{L \rightarrow \infty} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L)$$

proving (22a). ■

An intuitive interpretation of Theorem V.3 is that, as $L \rightarrow \infty$, the zero error capacity is the only part of the transmitted information rate that we may hope to preserve.

Another interpretation of this result can be obtained by considering the concept of *common information* as defined in [24]. As we discuss in more detail in Appendix III, the zero error capacity can be defined through common information in a similar manner as ordinary capacity can be defined through mutual information. Hence as $L \rightarrow \infty$, the only part of mutual information we can preserve between the input and the output of the cascade is exactly the common information between them.

The following examples illustrate the use of Theorem V.2 and Theorem V.3.

Example V.1. Consider $\mathbf{V} = \tilde{\mathbf{Q}}$, where $\tilde{\mathbf{Q}}$ is as in Example IV.1. The corresponding transition matrix \mathbf{V} and graph $G(\mathbf{V})$ are depicted in Figure 3.

For this channel [7]

$$C_0(\mathbf{V}) = M_0(\mathbf{V}) = \log 2,$$

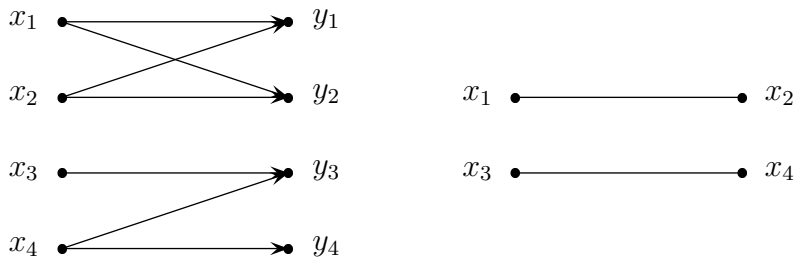


Fig. 3. The channel \mathbf{V} in Example V.1 and its graph $G(\mathbf{V})$.

that is, the zero error capacity is achieved by a zero error code of blocklength one (this code for example might use x_1 and x_3 that are non adjacent). Theorem V.3 states that for an infinite cascade of these channels, for any finite N

$$\lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) = \lim_{L \rightarrow \infty} C_{1,L}(\mathbf{V}) = \log 2.$$

In other words, if N is restricted to be finite, $N = 1$ is optimal.

The limiting capacity can be achieved by using at all intermediate nodes the decoder

$$\mathbf{M}_D = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}^T,$$

followed by the encoder

$$\mathbf{M}_E = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Notice that, for this specific example, the product $\mathbf{M}_D \mathbf{M}_E$ is actually equivalent to forwarding. In fact, a decoder can not distinguish between x_1 and x_2 , and between x_2 and x_3 . Nonetheless, x_1 or x_2 are never mistaken for x_3 or x_4 . Moreover, in the limit for large L , \mathbf{M}_E can be used at the source and \mathbf{M}_D can be used at the destination without loss of optimality. Hence, for long line networks of channels \mathbf{V} , the intermediate nodes need not to perform any processing at all, the source conveys the bits with value zero by sending, for example, x_1 or x_2 , and the bits with value one by sending x_3 or x_4 . The destination decodes a bit to be zero if either y_1 or y_2 is received, and one otherwise. This scheme does not incur any delay, does not require any intermediate processing, and is optimal as long as the length of intermediate processing is restricted to be finite while $L \rightarrow \infty$. \diamond

The simplest non-trivial DMC for which the zero error capacity is achieved by a zero error code of length bigger than one, is the so called “pentagon” channel [7], which we analyze in more detail in the next example.

Example V.2. The Pentagon Channel

Consider the “pentagon” channel whose transition matrix \mathbf{V} , for $p \in (0, 1)$, is

$$\mathbf{V} = \begin{pmatrix} 1-p & p & 0 & 0 & 0 \\ 0 & 1-p & p & 0 & 0 \\ 0 & 0 & 1-p & p & 0 \\ 0 & 0 & 0 & 1-p & p \\ p & 0 & 0 & 0 & 1-p \end{pmatrix}.$$

The corresponding graph $G(\mathbf{V})$ is depicted in Figure 4.

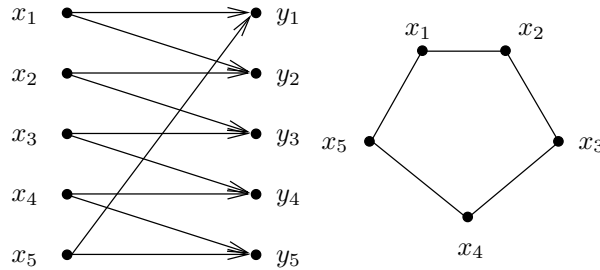


Fig. 4. The “pentagon” channel \mathbf{V} and its graph $G(\mathbf{V})$.

For this channel we have $M_0(\mathbf{V}) = 2$ (for example x_1 and x_3 are non adjacent) and $M_0(\mathbf{V}^{\otimes 2}) = 5$ (for example (x_1, x_1) , (x_2, x_3) , (x_3, x_5) , (x_4, x_2) , (x_5, x_4) are non adjacent). It was conjectured in [7] and shown in [25] that for this channel the zero error capacity is

$$C_0(\mathbf{V}) = \frac{1}{2}M_0(\mathbf{V}^{\otimes 2}) = \frac{1}{2}\log 5,$$

that is, a zero error code of blocklength two is optimal.

Theorem V.3 states that for an infinite cascade of “pentagon” channels

$$\lim_{L \rightarrow \infty} C_{1,L}(\mathbf{V}) = \log 2,$$

and

$$\lim_{L \rightarrow \infty} C_{2,L}(\mathbf{V}) = \frac{1}{2}\log 5.$$

Moreover, for any other finite N , we can never achieve more than $\frac{1}{2} \log 5$. Hence, for an infinite cascade of “pentagon” channels an intermediate processing of length $N = 2$ is optimal if N is restricted to be finite. The optimal limiting capacity can be achieved by using at all intermediate nodes and at the source a length-2 zero error encoder, and at all intermediate nodes and at the destination a length-2 zero error decoder. Notice that in this example, if the intermediate nodes simply forward the incoming data, then the limiting capacity is $\log D(\mathbf{V}) = \log 1 = 0$ (by Theorem IV.4), and this limit is approached exponentially fast with exponent $-2 \log |\lambda_2(\mathbf{V})| = -\log(1 - 2p(1 - p) \sin^2(2\pi/5))$ (by Theorem IV.5). In other words, intermediate processing is *necessary* if a non vanishing throughput is to be achieved in a long line network. Even a non-trivial one-symbol processing suffices to achieve a strictly positive end-to-end rate. \diamond

B. Cascade of Non-Identical Channels

A result similar to Theorem V.3 also holds if the channels to be cascaded are not identical. More formally, we consider the cascade of channels $\{\mathbf{V}_i\}_{i=1}^L$ where $\mathbf{V}_i \in \mathcal{V}$ for some set of stochastic matrices $\mathcal{V} \subset \bigcup_{i,j \leq n} \mathcal{S}_{i,j}$ with n being finite. As before, we allow intermediate nodes to perform some processing $\{\mathbf{M}_i\}_{i=1}^{L-1}$ of blocklength N and of appropriate dimensions, and we denote by $C_{N,L}(\{\mathbf{V}_i\})$ the capacity of the resulting cascade.

We now define a set

$$\mathcal{V}_{\text{i.o.}} \triangleq \{\mathbf{V} \in \mathcal{V} : M_0(\mathbf{V}^{\otimes N}) \text{ occurs i.o.}\}$$

that includes all matrices \mathbf{V} such that the corresponding $M_0(\mathbf{V}^{\otimes N})$ occurs infinitely often (i.o.) in the cascade. The following simple counting argument shows that the set $\mathcal{V}_{\text{i.o.}}$ is not empty. By definition the input alphabet size of all channels of the cascade is upper bounded by a finite number n . Then utilizing length- N processing implies that $M_0(\mathbf{V}_i^{\otimes N})$ can take a finite number of values, from 1 to n^N . But each of the matrices \mathbf{V}_i , $i = 1 \dots L$ with $L \rightarrow \infty$, will lead to one of these values. Therefore, at least one value will be observed an infinite number of times.

Theorem V.4.

$$\min_{\mathbf{V} \in \mathcal{V}} \frac{1}{N} \log M_0(\mathbf{V}^{\otimes N}) \leq \lim_{L \rightarrow \infty} C_{N,L}(\{\mathbf{V}_i\}) \leq \min_{\mathbf{V} \in \mathcal{V}_{\text{i.o.}}} \frac{1}{N} \log M_0(\mathbf{V}^{\otimes N})$$

Proof: We can always use a zero error code over each channel $\mathbf{V}_i^{\otimes N}$. The resulting capacity is $\min_{\mathbf{V} \in \mathcal{V}} \frac{1}{N} \log M_0(\mathbf{V}^{\otimes N})$, proving the lower bound.

For the upper bound, let $\widehat{\mathbf{V}} \in \mathcal{V}_{i.o.}$ be a matrix that achieves the minimum value. Replace every matrix \mathbf{V}_i in the cascade such that $M_0(\mathbf{V}_i^{\otimes N}) \neq M_0(\widehat{\mathbf{V}}^{\otimes N})$ by an arbitrary stochastic matrix $\widehat{\mathbf{M}}_i$. If we maximize capacity over all those $\{\widehat{\mathbf{M}}_i\}$ we obtain an upper bound on $C_{N,L}(\{\mathbf{V}_i\})$. Observe that with this last step we are really looking at a cascade of channels with identical zero error capacity for which we allow intermediate processing.

Consider now the remaining matrices $\mathbf{V}_i \in \mathcal{V}_{i.o.}$ with $M_0(\mathbf{V}_i^{\otimes N}) = M_0(\widehat{\mathbf{V}}^{\otimes N})$ of arbitrary dimension $j \times k$. Construct the matrix $\widetilde{\mathbf{W}}_i$ from $\mathbf{V}_i^{\otimes N}$ by duplicating the last row of $\mathbf{V}_i^{\otimes N}$ $n - j$ times and appending $n - k$ all zero columns. The resulting $\widetilde{\mathbf{W}}_i$ is a stochastic matrix of dimension $n \times n$. Replacing every matrix $\mathbf{V}_i^{\otimes N}$ by the corresponding $\widetilde{\mathbf{W}}_i$ and adapting the dimensions of the intermediate processings accordingly further upper bounds $C_{N,L}(\{\mathbf{V}_i\})$. Define $\mathbf{Q}_i \triangleq \widetilde{\mathbf{M}}_i \widetilde{\mathbf{W}}_i$, where $\widetilde{\mathbf{M}}_i$ is the optimal aggregate intermediate processing between channel $\widetilde{\mathbf{W}}_{i-1}$ and $\widetilde{\mathbf{W}}_i$ of this new cascade.

With this, the setup is now almost the same as in Theorem V.2 and by the same argument as in there we get that

$$\begin{aligned} \lim_{L \rightarrow \infty} C_{N,L}(\{\mathbf{V}_i\}) &\leq \lim_{L \rightarrow \infty} \frac{1}{N} C_{N,L}(\{\widetilde{\mathbf{W}}_i\}) \\ &= \lim_{L \rightarrow \infty} \frac{1}{N} C\left(\prod_{i=1}^L \mathbf{Q}_i\right) \\ &= \lim_{\ell \rightarrow \infty} \frac{1}{N} C(\mathbf{P} \widetilde{\mathbf{Q}}^\ell \widetilde{\mathbf{P}}) \\ &\leq \lim_{\ell \rightarrow \infty} \frac{1}{N} C(\widetilde{\mathbf{Q}}^\ell) \end{aligned}$$

for some stochastic matrices \mathbf{P} , $\widetilde{\mathbf{Q}}$, $\widetilde{\mathbf{P}}$. But from Theorem V.3,

$$\lim_{\ell \rightarrow \infty} \frac{1}{N} C(\widetilde{\mathbf{Q}}^\ell) = \frac{1}{N} M_0(\widetilde{\mathbf{Q}}),$$

and by Proposition V.1,

$$\frac{1}{N} M_0(\widetilde{\mathbf{Q}}) \leq \max_i \frac{1}{N} M_0(\widetilde{\mathbf{W}}_i) = \min_{\mathbf{v} \in \mathcal{V}_{i.o.}} \frac{1}{N} M_0(\mathbf{v}^{\otimes N}).$$

■

Notice that the upper and lower bound in Theorem V.4 coincide if $\mathcal{V} = \mathcal{V}_{i.o.}$. This is the case, for example, if every channel in the cascade appears an infinite number of times. A special case is when all $\{\mathbf{V}_i\}_{i=1}^L$ are identical, for which we recover the result of Theorem V.3. Following we give another example where the lower bound in Theorem V.4 is tight.

Example V.3. Consider a cascade consisting of $L - 1$ identical channels \mathbf{V} with $M_0(\mathbf{V}^{\otimes N}) > m$ for some integer m , and of one channel \mathbf{V}_0 such that

$$C(\mathbf{V}_0) = C_0(\mathbf{V}_0) = \log M_0(\mathbf{V}_0) = m,$$

i.e., for example the channel \mathbf{V}_0 is block diagonal with m blocks on the main diagonal, and each block is a rank one matrix. By the data processing inequality

$$C_{N,L}(\{\mathbf{V}_i\}) \leq \frac{1}{N} C(\mathbf{V}_0^{\otimes N}) = \log M_0(\mathbf{V}_0) = m = \min_{\mathbf{V} \in \mathcal{V}} \frac{1}{N} \log M_0(\mathbf{V}^{\otimes N}),$$

and hence the lower bound in Theorem V.4 is tight. \diamond

Notice that the problem of finding the limiting capacity of cascades of general arbitrary channels is as hard as finding the capacity for any finite cascade of channels. This can be seen from the following example. Let \mathbf{V} be the channel transition probability matrix of a BSC(p). Consider the cascade with $\mathbf{V}_1 = \mathbf{V}_2 = \mathbf{V}$ and $\mathbf{V}_i = \mathbf{I}$ for $i > 2$, then $\lim_{L \rightarrow \infty} C_{N,L}(\{\mathbf{V}_i\}) = C_{N,2}(\mathbf{V})$.

VI. BOUNDS ON CAPACITY

We will derive an upper and a lower bound on $C_{N,L}(\mathbf{V})$, the capacity of the cascade with optimal intermediate processing at the relays, that apply for all values of N and L .

A. Upper Bound

In this section we derive an upper bound for $C_{N,L}(\mathbf{V})$ expressed as a linear combination of the min-cut capacity and of a term reminiscent of the zero error capacity. The basic idea is to decompose the channel transition matrix $\mathbf{V}^{\otimes N}$ into a linear combination of two stochastic matrices, one of which with rank as close as possible to $M_0(\mathbf{V}^{\otimes N})$. We also discuss efficient algorithms to determine such a decomposition.

Theorem VI.1. *For any stochastic matrix \mathbf{V} and any integer N , if there exist two stochastic matrices \mathbf{A}_N and \mathbf{B}_N , and $\delta_N \in [0, 1]$ such that*

$$\mathbf{V}^{\otimes N} = \delta_N \mathbf{A}_N + (1 - \delta_N) \mathbf{B}_N \tag{25}$$

then

$$C_{N,L}(\mathbf{V}) \leq (1 - (1 - \delta_N)^{L-1}) \frac{\log \text{rank}(\mathbf{A}_N)}{N} + (1 - \delta_N)^{L-1} C(\mathbf{V}). \tag{26}$$

Remark. If $\log \text{rank}(\mathbf{A}_N)/N < C(\mathbf{V})$ and $\delta_N > 0$, the bound in (26) is strictly better than the min-cut upper bound $C_{N,L}(\mathbf{V}) \leq C(\mathbf{V})$.

Proof: Assume (25) holds, then,

$$C_{N,L}(\mathbf{V}) = \frac{1}{N} C\left(\mathbf{W} \prod_{i=1}^{L-1} (\mathbf{M}_i \mathbf{W})\right) \quad (27a)$$

$$= \frac{1}{N} C\left((\delta_N \mathbf{A}_N + (1 - \delta_N) \mathbf{B}_N) \prod_{i=1}^{L-1} (\mathbf{M}_i \mathbf{W})\right) \quad (27b)$$

$$\leq \delta_N \frac{1}{N} C\left(\mathbf{A}_N \prod_{i=1}^{L-1} (\mathbf{M}_i \mathbf{W})\right) + (1 - \delta_N) \frac{1}{N} C\left(\mathbf{B}_N \prod_{i=1}^{L-1} (\mathbf{M}_i \mathbf{W})\right) \quad (27c)$$

$$\leq \delta_N \frac{C(\mathbf{A}_N)}{N} + (1 - \delta_N) \frac{1}{N} C\left(\mathbf{W} \prod_{i=2}^{L-1} (\mathbf{M}_i \mathbf{W})\right) \quad (27d)$$

where (27c) follows from the convexity of mutual information in the channel matrix, and (27d) from the data processing inequality. By repeating the same argument, we get

$$C_{N,L}(\mathbf{V}) \leq (1 - (1 - \delta_N)^{L-1}) \frac{C(\mathbf{A}_N)}{N} + (1 - \delta_N)^{L-1} C(\mathbf{V}). \quad (28)$$

We can further upper bound $C(\mathbf{A}_N)$ in (28) with the logarithm of the rank of \mathbf{A}_N [22] to yield (26). ■

The following example illustrates the use of Theorem VI.1.

Example VI.1. Consider again $\mathbf{V} = \tilde{\mathbf{Q}}$, where $\tilde{\mathbf{Q}}$ is given in (10) in Example IV.1. Then, we can take

$$\delta_1 = \min\{2p, 2(1-p), t\},$$

and $\mathbf{A}_1 = \tilde{\mathbf{Q}}^\infty$, where $\tilde{\mathbf{Q}}^\infty$ is given in (11). In this case $\log \text{rank}(\mathbf{A}_1) = \log(2) = C_0(\mathbf{V})$ and the bound in (26) becomes

$$C_{N,L}(\mathbf{V}) \leq (1 - (1 - \delta_1^N))^{L-1} C_0(\mathbf{V}) + (1 - \delta_1^N)^{L-1} C(\mathbf{V}).$$

For this channel $\lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) = C_0(\mathbf{V})$ for any finite N , as already pointed out in Example V.1. Hence the decay of $C_{N,L}(\mathbf{V})$ to $C_0(\mathbf{V})$ as L increases is exponential in L with exponent lower bounded by

$$E_L(\mathbf{V}) \triangleq \lim_{L \rightarrow \infty} -\frac{1}{L} \log(C_{N,L}(\mathbf{V}) - C_0(\mathbf{V})) \geq -\log(1 - \delta_1^N).$$

For $N = 1$, $E_L(\mathbf{V}) \geq -\log(1 - \delta_1) = -\log \max\{|1 - 2p|, 1 - t\}$.

Since $C_{N,L}(\mathbf{V}) \geq C(\mathbf{V}^L)$, we can upper bound the exponent of the rate of decay by the exponent of the rate of decay of $C(\mathbf{V}^L)$ to $C_0(\mathbf{V})$ derived in Example IV.2, to get

$$E_L(\mathbf{V}) \leq -\log \max\{(1 - 2p)^2, 1 - t\}.$$

By comparing the upper bound to $E_L(\mathbf{V})$ with the lower bound for $N = 1$, we see that the two bounds coincide for $|1 - 2p| < 1 - t$, thus giving the correct exponent for $C_{N,L}(\mathbf{V})$ in this regime. \diamond

In order to obtain the best bound for any given N , \mathbf{A}_N should be chosen to have the *smallest rank possible*. A possible choice is to take $\mathbf{A}_N = \mathbf{A}_1^{\otimes N}$ and $\delta_N = \delta_1^N$. With this the bound in (26) reduces to

$$C_{N,L}(\mathbf{V}) \leq (1 - (1 - \delta_1^N)^{L-1}) \log \text{rank}(\mathbf{A}_1) + (1 - \delta_1^N)^{L-1} C(\mathbf{V})$$

since $\text{rank}(\mathbf{A}_N) = (\text{rank}(\mathbf{A}_1))^N$. However, the choice $\mathbf{A}_N = \mathbf{A}_1^{\otimes N}$ does not give the best possible bound for $C_{N,L}(\mathbf{V})$ in (26) in general, as we show in Example VI.2.

Example VI.2. Consider again the ‘‘pentagon’’ channel introduced in Example V.2. For $N = 1$ we can find a matrix \mathbf{A}_1 with $\text{rank}(\mathbf{A}_1) = 3$. However, for $N = 2$ we can find a matrix \mathbf{A}_2 with $\text{rank}(\mathbf{A}_2) = 8 < \text{rank}(\mathbf{A}_1)^2 = 9$. \diamond

Note that for any matrix \mathbf{A}_N such that (25) holds we have

$$\frac{1}{N} \log M_0(\mathbf{V}^{\otimes N}) = \lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) \leq \frac{\log \text{rank}(\mathbf{A}_N)}{N}. \quad (29)$$

If for some N we find $\text{rank}(\mathbf{A}_N) = M_0(\mathbf{V}^{\otimes N})$ (like in Example VI.1), then the bound in (26), is strictly between $C_{\infty,L}(\mathbf{V}) = C(\mathbf{V})$ and $C_{N,\infty}(\mathbf{V}) = 1/N \log M_0(\mathbf{V}^{\otimes N})$. In this case, the decay of $C_{N,L}(\mathbf{V})$ to $C_{N,\infty}(\mathbf{V})$ is exponentially fast in L . We have already seen in Section IV-C that if we impose the constraint that all $\{\mathbf{M}_i\}_{i=1}^{L-1}$ are identical then the limiting capacity is achieved exponentially fast in L , and we gave tight upper and lower bounds on the exponent. If $\text{rank}(\mathbf{A}_N) = M_0(\mathbf{V}^{\otimes N})$ for some N , then exponential decay also applies to non identical processing. The exponent that can be derived from (26), namely $E_L \geq -\log(1 - \delta_N)$, is however not tight in general.

The problem of finding the matrix \mathbf{A}_N with minimum rank is equivalent to the *Set Cover Problem* described as follows [26]. Given a universe \mathcal{U} of n elements, a collection $\mathcal{S} = \{\mathcal{S}_1 \dots \mathcal{S}_m\}$ of subsets of \mathcal{U} , and a cost function for each subset in \mathcal{S} , find a minimum cost subcollection

of \mathcal{S} that covers all the elements in \mathcal{U} . This problem can be formulated as an integer program as follows. Assign a variable x_i for each set $\mathcal{S}_i \in \mathcal{S}$, where $x_i = 1$ if set \mathcal{S}_i takes part in the subcollection and $x_i = 0$ otherwise. The constraint is that every elements in \mathcal{U} must belong to at least one of the picked sets \mathcal{S}_i . The set cover problem and its Linear Program (LP) relaxation (Primal) are provided in Table I, for the special case where the cost of all sets is one, which is the case of interest here. A variety of approximation algorithms are available in the literature for the set cover problem [26]. Those algorithms run in polynomial time and provide approximations with gap at most $\log n$ from the optimal solution.

In our case, the universe \mathcal{U} is the set of $n = |\mathcal{X}|^N$ inputs of the channel $\mathbf{V}^{\otimes N}$. The set of $m = |\mathcal{Y}|^N$ outputs defines \mathcal{S} , in that the subset \mathcal{S}_i contains the inputs that result with nonzero probability in output i , for $i = 1, \dots, m$.

The solution of the set cover problem in our case provides us with a set of outputs that cover all inputs. Indeed, let OPT_{A_N} denote the minimum cost of the set cover problem. We can construct a matrix \mathbf{A}_N that has rank OPT_{A_N} as follows. Take the binary matrix $\lceil \mathbf{V}^{\otimes N} \rceil$, where the ceiling operation is component-wise. To construct matrix \mathbf{A}_N , replace with the all-zero columns those columns of $\lceil \mathbf{V}^{\otimes N} \rceil$ that correspond to outputs where $x_i = 0$ in the set cover problem. Then normalize all rows so that they sum to one. The matrix \mathbf{A}_N contains OPT_{A_N} nonzero columns and $m - \text{OPT}_{A_N}$ all-zero columns, and thus $\text{rank}(\mathbf{A}_N) = \text{OPT}_{A_N}$. Moreover, \mathbf{A}_N is a valid stochastic matrix because its nonzero columns “cover” all inputs, i.e., each row has at least one nonzero element, and all rows sum to one. Choose δ_N to be the largest number such that the matrix

$$\mathbf{B}_N = \frac{1}{1 - \delta_N} (\mathbf{V}^{\otimes N} - \delta_N \mathbf{A}_N)$$

is a valid stochastic matrix. It is easy to see that this δ_N is at least as large as the minimum of the entries of $\mathbf{V}^{\otimes N}$ that correspond to a non-zero entry in \mathbf{A}_N . Moreover, by using one of the available approximation algorithms, we can calculate in polynomial time a matrix \mathbf{A}_N that has rank bigger than the minimum by a factor of at most $\log n = N \log |\mathcal{X}|$. As we are really interested in $\log(\text{rank}(\mathbf{A}_N))/N$ this implies that the loss we incur by using these approximation algorithms to find a \mathbf{A}_N matrix goes to zero as $N \rightarrow \infty$.

The problem of finding the matrix \mathbf{A}_N with minimum rank is closely related to computing $M_0(\mathbf{V}^{\otimes N})$. To see this, consider the (strong) LP dual of the set cover LP relaxation described

TABLE I

MAXIMUM INDEPENDENT SET PROBLEM, SET COVER PROBLEM, AND THEIR LP RELAXATIONS.

Set Cover Problem $\min \sum_i x_i$ $\sum_{i:u \in \mathcal{S}_i} x_i \geq 1, \forall u \in \mathcal{U}$ $x_i \in \{0, 1\}, \forall i$	LP Relaxation (Primal) $\min \sum_i x_i$ $\sum_{i:u \in \mathcal{S}_i} x_i \geq 1, \forall u \in \mathcal{U}$ $0 \leq x_i \leq 1, \forall i$
Max Independent Set Problem $\max \sum_j y_j$ $\sum_{j \in \mathcal{S}_i} y_j \leq 1 \forall \mathcal{S}_i \in \mathcal{S}$ $y_j \in \{0, 1\}, \forall j$	LP Relaxation (Dual) $\max \sum_j y_j$ $\sum_{j \in \mathcal{S}_i} y_j \leq 1 \forall \mathcal{S}_i \in \mathcal{S}$ $0 \leq y_j \leq 1, \forall j$

in Table I. The dual LP is the LP relaxation of the *Maximum Independent Set Problem*. The maximum independent set problem takes as input a graph adjacency matrix and calculates the graph independence number. We already pointed out in Section II-A that $M_0(\mathbf{V}^{\otimes N})$ is the independence number of the graph $G(\mathbf{V}^{\otimes N})$. This maximum independent set problem can be formulated as an integer program as follows. Assign a variable y_j for each vertex of the graph, $y_j = 1$ if the vertex takes part in the independent set and $y_j = 0$ otherwise. The constraint is that no two picked vertices are connected with an edge.

For our purposes we use the adjacency matrix corresponding to the graph $G(\mathbf{V}^{\otimes N})$. The solution of the maximum independent set problem for $G(\mathbf{V}^{\otimes N})$ directly leads to the construction of an optimal zero-error code of length N for the channel \mathbf{V} . Denote by OPT_{M_0} this optimal solution. Obviously, OPT_{A_N} is lower-bounded by OPT_{M_0} . In fact, OPT_{A_N} is the minimum number of outputs such that all inputs are covered. Since all inputs are covered, this implies that any $\text{OPT}_{A_N} + 1$ inputs have at least one output in common. Thus $\text{OPT}_{M_0} \leq \text{OPT}_{A_N}$.

Note that, in the instances where the integrality gap between the maximum independent set problem and its LP relaxation is small, then there exists a matrix \mathbf{A}_N such that $\text{rank}(\mathbf{A}_N) \approx M_0(\mathbf{V}^{\otimes N})$, i.e., the inequality in (29) becomes an equality.

B. Lower Bound

We next derive a lower bound on $C_{N,L}(\mathbf{V})$. To do so, we choose a particular (possibly suboptimal) communication scheme and find a lower bound on the rate achievable with this scheme. Assume we use an inner encoder \mathcal{M}_E at the source A_0 and a corresponding maximum

likelihood decoder M_D at the relay A_1 . At A_1 the message is then re-encoded using again M_E and transmitted over the second channel. We continue in the same manner at every A_i , and the destination A_L uses an inner decoder M_D . This corresponds to using intermediate processing $M_i = M = M_D M_E$ for all i . The rate r of this inner code is determined by ρ , the rank of M , through

$$\rho = \exp(Nr).$$

Note that this scheme constructs an overall channel

$$\mathbf{W}_{\text{eq}} = (\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L = \mathbf{Q}^L$$

with $\mathbf{Q} \triangleq \mathbf{M}_E \mathbf{W} \mathbf{M}_D \in \mathcal{S}_{\rho, \rho}$. The source A_0 and the destination A_L will then use an outer code over the channel \mathbf{Q}^L .

Using random coding arguments [17], we know that there exist good codes (defined by the tuple $(\mathbf{M}_E, \mathbf{M}_D)$) in the sense that the average probability of decoding error is bounded by

$$P_e(r) \leq \exp(-NE_r(r)) \triangleq \delta(r),$$

where $E_r(\cdot)$ is the random coding error exponent for the channel \mathbf{V} as a function of the rate r . We use such a good code as our inner code. With this, we know that the trace of \mathbf{Q} is lower bounded by $\rho(1 - \delta(r))$, but unfortunately this results gives no information about the off-diagonal entries of \mathbf{Q} . To get a lower bound on the achievable rate, we will construct the worst cascade \mathbf{Q}^L such that \mathbf{Q} satisfies the trace constraint. This worst cascade is found in Appendix Lemma II.1, whose proof is in Appendix II. The resulting lower bound for $C_{N,L}(\mathbf{V})$ is given in the next theorem.

Theorem VI.2.

$$C_{N,L}(\mathbf{V}) \geq \max_{r \in [0, C(\mathbf{V})]} r(1 - \delta(r))^L - \frac{1}{N}, \quad (30)$$

and for $N \rightarrow \infty$ the bound is tight, i.e.,

$$\lim_{N \rightarrow \infty} \max_{r \in [0, C(\mathbf{V})]} r(1 - \delta(r))^L - \frac{1}{N} = C(\mathbf{V}). \quad (31)$$

Proof: Let $\mathcal{W}_1(a)$ be the set of stochastic matrices \mathbf{Q}^L such that $\text{tr}(\mathbf{Q}) \geq \rho a$, i.e.,

$$\mathcal{W}_1(a) \triangleq \{\widetilde{\mathbf{W}} : \widetilde{\mathbf{W}} = \mathbf{Q}^L, \mathbf{Q} \in \mathcal{S}_{\rho, \rho}, \text{tr}(\mathbf{Q}) \geq \rho a\}$$

and $\mathcal{W}_2(a)$ be the set of stochastic matrices \mathbf{Q} such that $\text{tr}(\mathbf{Q}) \geq \rho a^L$, i.e.,

$$\mathcal{W}_2(a) \triangleq \{\widetilde{\mathbf{W}} : \widetilde{\mathbf{W}} \in \mathcal{S}_{\rho,\rho}, \text{tr}(\widetilde{\mathbf{W}}) \geq \rho a^L\}.$$

For any $\mathbf{Q} \in \mathcal{S}_{\rho,\rho}$ with $\text{tr}(\mathbf{Q}) \geq \rho a$

$$\text{tr}(\mathbf{Q}^L) = \sum_{i_1, \dots, i_{L-1}, j} q_{ji_1} q_{i_1 i_2} \cdots q_{i_{L-1} j} \geq \sum_j q_{jj}^L \geq \rho a^L,$$

where $q_{ji} = [\mathbf{Q}]_{j,i}$. Hence $\mathcal{W}_1(a) \subset \mathcal{W}_2(a)$ and therefore

$$\begin{aligned} C_{N,L}(V) &\geq \frac{1}{N} \inf_{\widetilde{\mathbf{W}} \in \mathcal{W}_1(1-P_e(r))} \max_{\mathbf{p}} I(\mathbf{p}, \widetilde{\mathbf{W}}) \\ &\geq \frac{1}{N} \inf_{\widetilde{\mathbf{W}} \in \mathcal{W}_2(1-P_e(r))} \max_{\mathbf{p}} I(\mathbf{p}, \widetilde{\mathbf{W}}) \\ &\geq \frac{1}{N} \max_{\mathbf{p}} \min_{\widetilde{\mathbf{W}} \in \mathcal{W}_2(1-P_e(r))} I(\mathbf{p}, \widetilde{\mathbf{W}}). \end{aligned} \quad (32)$$

The minimization in (32) can be solved using Appendix Lemma II.1. If $(1 - P_e(r))^L \leq 1/\rho$ the worst channel is $\frac{1}{\rho} \mathbf{1}\mathbf{1}^T$, and capacity is (trivially) lower bounded by 0. Else the worst channel has diagonal entries equal to $(1 - P_e(r))^L$ and all other entries equal to $(1 - (1 - P_e(r))^L)/(\rho - 1)$. In both cases the minimizer is a symmetric channel [1]. For symmetric channels the capacity is given by

$$\log(\rho) - H(\tilde{\mathbf{w}}_1),$$

where $\tilde{\mathbf{w}}_1$ is the first row of the minimizing channel matrix $\widetilde{\mathbf{W}}$. Thus the expression in (32) can be calculated as

$$\begin{aligned} &\frac{1}{N} \left(\log \rho - (1 - (1 - P_e(r))^L) \log(\rho - 1) - \mathcal{H}((1 - P_e(r))^L) \right) \\ &\geq r(1 - P_e(r))^L - \frac{1}{N} \\ &\geq r(1 - \delta(r))^L - \frac{1}{N}, \end{aligned}$$

where the above bound is valid for all values of the inner code rate r such that

$$(1 - P_e(r))^L \geq (1 - \delta(r))^L \geq 1/\rho.$$

If $(1 - \delta(r))^L \leq 1/\rho$ the bound becomes

$$\frac{\log(\rho)}{N} (1 - \delta(r))^L - \frac{1}{N} \leq \frac{1}{N} \left(\frac{\log(\rho)}{\rho} - 1 \right) \leq 0.$$

Hence the bound is also (trivially) valid in this range, and we can maximize over $r \in [0, C(\mathbf{V})]$, thus showing that

$$C_{N,L}(\mathbf{V}) \geq \sup_{r \in [0, C(\mathbf{V})]} r(1 - P_e(r))^L - \frac{1}{N}.$$

As $E_r(r)$ is continuous in r [17], $r(1 - \delta(r))^L$ is continuous in r as well. Moreover, the set of r over which we optimize is compact, and hence we can replace the sup with a max, proving (30).

We will now show that the bound is tight as $N \rightarrow \infty$. As $E_r(r)$ is strictly positive for all $r < C(\mathbf{V})$ [17], we have for every $\varepsilon \geq 0$

$$\lim_{N \rightarrow \infty} (C(\mathbf{V}) - \varepsilon) \left(1 - \exp(-NE_r(C(\mathbf{V}) - \varepsilon)) \right)^L - \frac{1}{N} = C(\mathbf{V}) - \varepsilon.$$

As ε is arbitrary (31) follows. ■

Example VI.3. Consider a cascade of $L = 2$ BSC(p). Figure 5 compares for this channel the lower bound derived in this section with the forwarding capacity $C_{1,2}(\mathbf{V})$ (i.e., $\mathbf{M} = \mathbf{I}$) and the min-cut capacity $C(\mathbf{V})$ (achievable when $N \rightarrow \infty$). It can be seen that, while the lower bound is not very good for small values of N , it is quite good for larger values of N and tight as $N \rightarrow \infty$.

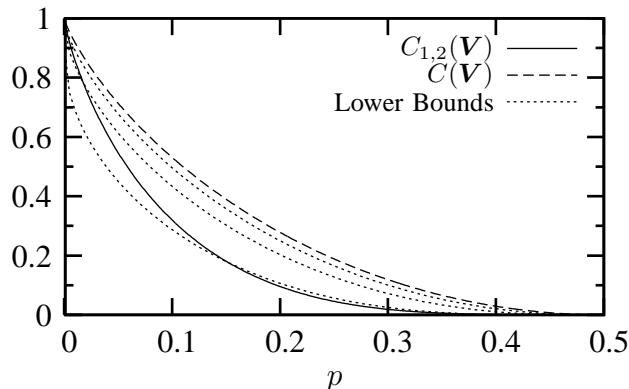


Fig. 5. Lower bounds on capacity for different values of blocklength $N \in \{10^2, 10^3, 10^4\}$ (from bottom to top in the figure). Also shown are the min-cut upper bound $C(\mathbf{V})$ and the forwarding lower bound $C_{1,2}(\mathbf{V})$. Note that the lower bound derived in this section is not very good for small values of N . Indeed for $N = 100$ the forwarding lower bound yields better results for some values of p . The bound is, however, tight for $N \rightarrow \infty$ and is equal to the min-cut capacity in this case.

How fast this convergence of the lower bound to the limiting expression and upper bound $C(\mathbf{V})$ takes place as a function of N is depicted in Figure 6 for various values of crossover

probability p . It can be seen that the lower bound is already quite close to the limiting expression for $N \geq 10^4$. \diamond

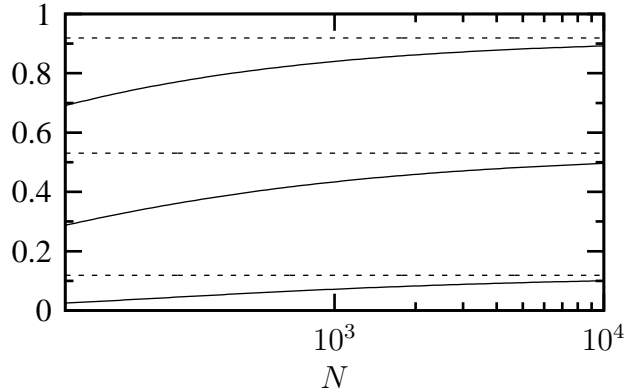


Fig. 6. Lower bounds on capacity for different values of crossover probability $p \in \{0.01, 0.1, 0.3\}$ (from top to bottom in the figure, solid lines) as a function of blocklength N of intermediate processing. Also shown are the corresponding limiting expressions $C(\mathbf{V})$ (dashed lines).

VII. SCALING LAWS

In this section, we investigate how the blocklength N needs to scale with the network length L in order to achieve a constant fraction of the min-cut capacity, as opposed to the zero error capacity. We will show that logarithmic growth of N with L is sufficient, and in many cases also necessary.

Any rate strictly below the zero error capacity can be achieved with finite blocklength processing, because the limit superior in the definition of zero error capacity in (2) is a true limit [16]. Here we are interested in rates that indeed need an infinite blocklength. Hence, for any $\alpha \in [0, 1]$ we define

$$N^*(L, \alpha) \triangleq \inf \{N : C_{N,L}(\mathbf{V}) \geq (1 - \alpha)C_0(\mathbf{V}) + \alpha C(\mathbf{V}) \triangleq R(\alpha)\}.$$

The next theorem gives an upper bound on $N^*(L, \alpha)$, establishing that logarithmic growth of N with L is sufficient to achieve $R(\alpha)$.

Theorem VII.1. *For every $\varepsilon > 0$ there is a N_0 such that for all $N \geq N_0$*

$$N^*(L, \alpha) \leq \min_{r \in [R(\alpha) + \varepsilon, C(\mathbf{V})]} \frac{1}{E_r(r)} \log \left(\frac{L}{1 - \frac{R(\alpha) + \varepsilon}{r}} \right) \quad (33)$$

is sufficient to achieve $R(\alpha)$.

Proof: Theorem VI.2 asserts that for any $r \in [0, C(\mathbf{V})]$

$$C_{N,L}(\mathbf{V}) \geq r \left(1 - \exp(-N E_r(r)) \right)^L - \frac{1}{N}. \quad (34)$$

Since the right hand side of (34) is always smaller than r , in order to attain $R(\alpha)$, r must satisfy $r \geq R(\alpha)$. By setting the right hand side of (34) equal to $R(\alpha)$, and by using the fact that $(1-x)^L \geq 1-Lx$ for all $x \in [0, 1]$, we get

$$\log \left(1 - \frac{R(\alpha) + \frac{1}{N}}{r} \right) \leq \log(L) - N E_r(r).$$

Hence, for every $\varepsilon > 0$ there exists and N_0 such that for all $N > N_0$

$$N \leq \frac{1}{E_r(r)} \left(\log(L) - \log \left(1 - \frac{R(\alpha) + \varepsilon}{r} \right) \right). \quad (35)$$

But since this is true for all $r \in [R(\alpha) + \varepsilon, C(\mathbf{V})]$, we can minimize the right hand side of (35) over r to get the tightest bound. Since the right hand side of (35) is a continuous function of r and the optimization set is compact, the result in (33) follows. ■

Notice that by solving (34) for L , we get also that for fixed N to guarantee a rate R we can cascade at most

$$L \leq \max_{r \in [R + \frac{1}{N}, C(\mathbf{V})]} \frac{1}{P_e(r)} \left(1 - \frac{R + \frac{1}{N}}{r} \right)$$

channels. This last result is interesting as it recovers, as a special case, a scaling law for the Z channel derived in [12].

The next theorem establishes that logarithmic growth of N with L is necessary to achieve $R(\alpha)$ for all $\alpha \geq \beta_m$ where β_m is a non-negative constant.

Theorem VII.2.

$$N^*(L, \alpha) \geq \frac{\log(L-1) - \log \log \frac{1}{\alpha - \beta_m}}{\frac{1}{m} \log \frac{1}{\delta_m}} \quad (36)$$

is necessary to achieve $R(\alpha)$ for all

$$\alpha \geq \beta_m \triangleq \frac{\frac{1}{m} \log \text{rank}(\mathbf{A}_m) - C_0(\mathbf{V})}{C(\mathbf{V}) - C_0(\mathbf{V})},$$

where m is any integer such that the matrices $\mathbf{A}_m, \mathbf{B}_m \in \mathcal{S}_{|\mathcal{X}|^m, |\mathcal{Y}|^m}$, and the real-valued constant $\delta_m \in [0, 1]$ in the decomposition

$$\mathbf{V}^{\otimes m} = \delta_m \mathbf{A}_m + (1 - \delta_m) \mathbf{B}_m$$

satisfies $C_0(\mathbf{V}) \leq \log \text{rank}(\mathbf{A}_m)/m \leq C(\mathbf{V})$ and $\delta_m > 0$.

Proof: From Theorem VI.1,

$$C_{N,L}(\mathbf{V}) \leq (1 - (1 - \delta_m^{N/m})^{L-1}) \frac{1}{m} \log \text{rank}(\mathbf{A}_m) + (1 - \delta_m^{N/m})^{L-1} C(\mathbf{V}).$$

In order to achieve $R(\alpha) = (1 - \alpha)C_0(\mathbf{V}) + \alpha C(\mathbf{V})$ it is hence necessary that

$$(1 - \alpha)C_0(\mathbf{V}) + \alpha C(\mathbf{V}) \leq (1 - (1 - \delta_m^{N/m})^{L-1}) \frac{1}{m} \log \text{rank}(\mathbf{A}_m) + (1 - \delta_m^{N/m})^{L-1} C(\mathbf{V}).$$

By re-arranging the different terms, we get

$$\alpha - \underbrace{\frac{\frac{1}{m} \log \text{rank}(\mathbf{A}_m) - C_0(\mathbf{V})}{C(\mathbf{V}) - C_0(\mathbf{V})}}_{\triangleq \beta_m} \leq (1 - \delta_m^{N/m})^{L-1} \underbrace{\frac{C(\mathbf{V}) - \frac{1}{m} \log \text{rank}(\mathbf{A}_m)}{C(\mathbf{V}) - C_0(\mathbf{V})}}_{\triangleq \gamma_m}.$$

If the value of m is such that the hypothesis of the theorem are satisfied we have that $\beta_m \in [0, 1]$ and $\gamma_m \in [0, 1]$, and hence

$$(1 - \delta_m^{N/m})^{L-1} \geq \alpha - \beta_m.$$

By using the fact that $-\log(1 - x) \geq x$ for all $x \in [0, 1]$, we get that for all $\alpha \geq \beta_m$

$$\log \frac{1}{\alpha - \beta_m} \geq (L - 1) \log \frac{1}{1 - \delta_m^{N/m}} \geq (L - 1) \delta_m^{N/m},$$

and hence

$$N \frac{1}{m} \log \frac{1}{\delta_m} \geq \log(L - 1) - \log \log \frac{1}{\alpha - \beta_m},$$

from which the result in (36) follows. ■

Notice that, since we are interested in the regime $N \gg 1$, the assumptions of the theorem about the integer m can be relaxed to $\lim_{m \rightarrow \infty} 1/m \log \text{rank}(\mathbf{A}_m) \leq C(\mathbf{V})$ and $\limsup_{m \rightarrow \infty} -1/m \log \delta_m > 0$. The inequality in (29) already implies that $C_0(\mathbf{V}) \leq \lim_{m \rightarrow \infty} 1/m \log \text{rank}(\mathbf{A}_m)$.

Example VII.1. Consider again the ‘‘pentagon’’ channel introduced in Example V.2. In Example VI.2 we showed that a decomposition with $\log \text{rank}(\mathbf{A}_2) = 8$ exists (here $C(\mathbf{V}) = \log 5$ and $C_0(\mathbf{V}) = 1/2 \log 5$). Hence

$$\beta_2 = \frac{\frac{1}{2} \log \text{rank}(\mathbf{A}_2) - C_0(\mathbf{V})}{C(\mathbf{V}) - C_0(\mathbf{V})} = \frac{\log 8/5}{\log 5} = 0.292,$$

showing that logarithmic growth of N with L is necessary to achieve any $R(\alpha)$ with $\alpha \geq 0.292$.

With $m = 2$, Theorem VII.2 does in this case not allow to state that logarithmic growth is necessary in the range $\alpha < 0.292$.

Example VII.2. Scaling Laws for Line Networks of Binary Symmetric Channels

Consider a cascade of L BSC(p) as in Example II.1. We know that any finite length processing performed at every node in the network will result in a zero end-to-end rate as $L \rightarrow \infty$. For example, if all the nodes use the encoder \mathbf{M}_E followed by the decoder \mathbf{M}_D of a length- N repetition code, including the source and the destination, the original network of BSCs is equivalent to a network of DMCs with transition probability matrix $\mathbf{M}_E \mathbf{V}^{\otimes N} \mathbf{M}_D$ corresponding to a BSC(p_N). The eigenvalues of a BSC(p_N) transition probability matrix are $\{1, 1 - 2p_N\}$. Hence, since there is only one eigenvalue of maximum modulus, the end-to-end achievable rate tends to $\log D(\mathbf{M}_E \mathbf{V}^{\otimes N} \mathbf{M}_D) = \log 1 = 0$. This limiting value is reached exponentially fast for any fixed finite N , with exponent depending on $1 - 2p_N$, i.e.,

$$\frac{1}{N} C \left(\frac{1 - (1 - 2p_N)^L}{2} \right) \approx \frac{1}{2N} (1 - 2p_N)^{2L} = \frac{1}{2N} \exp(L 2 \log |1 - 2p_N|) \rightarrow 0 \quad \text{as } L \rightarrow \infty$$

From Theorem VII.1, we see that logarithmic growth of N with L is *sufficient* to achieve any fraction of the min-cut capacity. We will now show that $N = \Theta(\log L)$ but also *necessary* to achieve any positive fraction of the min-cut capacity. We proceed as follows. The equivalent channel matrix between any pairs of nodes is $\mathbf{V}^{\otimes N}$ whose smallest entry is p^N . By collecting p^N from all the entries of $\mathbf{V}^{\otimes N}$, we can write

$$\mathbf{V}^{\otimes N} = p^N \mathbf{1}\mathbf{1}^T + (1 - (2p)^N) \mathbf{B},$$

for some stochastic matrix $\mathbf{B} \in \mathcal{S}_{2^N, 2^N}$, and where $\mathbf{1}$ is the all one column vector of length 2^N . By using this decomposition we get from Theorem VII.2 that

$$N^*(L, \alpha) \geq \frac{(L - 1) - \log \log \frac{1}{\alpha}}{\log \frac{1}{\delta}},$$

with $\delta \triangleq 2p$. Hence logarithmic growth of N with L is necessary for all positive rates. \diamond

VIII. CONCLUSIONS

In this paper, we investigated communication through a cascade of L channels, where intermediate nodes can perform a processing of finite complexity, where complexity is measured by the processing length N .

When N is fixed, L goes to infinity, and all relays use the same processing, we showed that the limiting capacity can be easily computed as the logarithm of the number of eigenvalues of

maximum modulus of the equivalent channel transition probability matrix that comprises the intermediate processing as part of the channel. We also showed how the rate of decay to the limiting capacity is related to the second largest eigenvalue modulus of the equivalent channel transition probability matrix.

When N is fixed and L goes to infinity, we showed that the optimal finite complexity processing is identical at each relay and corresponds to the use of an optimal zero error code of blocklength N for the underlying channel. The resulting limiting capacity is then the rate of this zero error code and can never exceed the zero error capacity of the underlying channel.

We also derived bounds on the capacity of finite length cascades and used them to show that logarithmic growth of N with L is sufficient to achieve any constant fraction of the min-cut capacity above the zero error capacity. Moreover, we showed that for rates above some threshold, logarithmic growth is also necessary. We conjecture that logarithmic growth is in fact necessary to achieve any rate above the zero error capacity. To prove our conjecture, a tighter upper bound on capacity is needed.

The fact that for a fixed intermediate processing the decay to the limiting capacity is exponential contrasts the logarithmic scaling law found for the optimal N as a function of L and emphasizes the importance of a well chosen intermediate processing at the relays.

In this work, we did not address the problem of identifying the optimal processing for any finite pair (N, L) , that is very interesting but combinatorial in nature. We view this work as a first step towards a more comprehensive understanding of how we should efficiently use limited network resources to achieve reliable communication. In our setting, resources are constrained in terms of complexity of intermediate node processing. Extending this work for more general (other than line) networks, and more general traffic configurations and resource constraints is part of ongoing work.

ACKNOWLEDGMENT

The authors would like to thank J. Körner for pointing out the connection with common information, and C. Chekuri, O. Lévêque, B. Rimoldi, E. Telatar and D. Shah for many useful discussions.

APPENDIX I
PROOF OF THEOREM IV.5

The proof of Theorem IV.5 will be broken up in several lemmas. The first lemma proves the lower bound in Theorem IV.5.

Appendix Lemma I.1. *For any stochastic matrix \mathbf{Q}*

$$E_L(\mathbf{Q}) \geq -\log |\lambda_2(\mathbf{Q})|.$$

Proof: Assume, without loss of generality, that \mathbf{Q} is in canonical form with diagonal irreducible submatrices $\{\mathbf{P}_i\}$ with periods $\{d_i\}$. Call d the least common multiple of the $\{d_i\}$. Equation (15) asserts that we can consider powers of \mathbf{Q} which are a multiple of d , that is,

$$E_L(\mathbf{Q}) = \liminf_{L \rightarrow \infty} -\frac{1}{dL} \log (C(\mathbf{Q}^{dL}) - C(\mathbf{Q}^\infty)).$$

From Lemma IV.1

$$\mathbf{Q}^{dL} = \mathbf{Q}^\infty + \mathbf{E}^{(dL)},$$

where, from Lemma IV.2, the components of $\mathbf{E}^{(dL)}$ converge to 0 exponentially fast in L with exponent no smaller than

$$-\log |\lambda_2(\mathbf{Q}^d)| = -d \log |\lambda_2(\mathbf{Q})|. \quad (37)$$

We will now use an upper bound on capacity from [27], which states that

$$C(\mathbf{Q}) \leq \log \sum_{\mathbf{y}} \max_{\mathbf{x}} q(\mathbf{y}|\mathbf{x}).$$

It is easily seen that the maximum entry of every column of \mathbf{Q}^∞ lies in the corresponding diagonal block. Moreover, the exponents of the entries in $\mathbf{E}^{(dL)}$ are all lower bounded by (37). Thus, for some polynomial $b(L)$,

$$C(\mathbf{Q}^{dL}) \leq \log \left(D(\mathbf{Q}) + b(L) \exp(dL \log |\lambda_2(\mathbf{Q})|) \right),$$

where $D(\mathbf{Q})$ is the number of primitive diagonal blocks of \mathbf{Q}^d , and where we used the fact that the stationary distribution of each such diagonal block sums to one. Recall that $D(\mathbf{Q})$ is also the

number of eigenvalues of modulus one of the matrix \mathbf{Q} (and \mathbf{Q}^d) and that, from Theorem IV.4, $C(\mathbf{Q}^\infty) = \log D(\mathbf{Q})$. Then

$$\begin{aligned} & \log \left(\log \left(D(\mathbf{Q}) + b(L) \exp(dL \log |\lambda_2(\mathbf{Q})|) \right) - \log(D(\mathbf{Q})) \right) \\ &= \log \left(\log \left(1 + \tilde{b}(L) \exp(dL \log |\lambda_2(\mathbf{Q})|) \right) \right) \\ &\leq \log \left(\tilde{b}(L) \exp(dL \log |\lambda_2(\mathbf{Q})|) \right), \end{aligned}$$

where $\tilde{b}(L) \triangleq b(L)/D(\mathbf{Q})$ and we used the expansion of $\log(x)$. Thus we get that

$$E_L(\mathbf{Q}) \geq -\log |\lambda_2(\mathbf{Q})|$$

as $\liminf_{L \rightarrow \infty} -\frac{1}{dL} \log \tilde{b}(L) = 0$. ■

As a next step, we will restrict attention to primitive matrices \mathbf{Q} , for which we can find the exponent $E_L(\mathbf{Q})$ exactly.

Appendix Lemma I.2. *If \mathbf{Q} is a primitive stochastic matrix then*

$$E_L(\mathbf{Q}) = -2 \log |\lambda_2(\mathbf{Q})|.$$

Proof: We will first show that $E_L(\mathbf{Q}) \geq -2 \log |\lambda_2(\mathbf{Q})|$. From Lemma IV.1 and IV.2

$$\mathbf{Q}^L = \mathbf{1}\boldsymbol{\pi} + \mathbf{E}^{(L)}, \tag{38}$$

where the components of $\mathbf{E}^{(L)}$ converge to 0 exponentially fast with exponent $-\log |\lambda_2(\mathbf{Q})|$ and where the row vector $\boldsymbol{\pi}$ is the (unique) stationary distribution of \mathbf{Q} . Define the set

$$\mathcal{A}(\beta) \triangleq \{\tilde{\mathbf{Q}} : \tilde{\mathbf{Q}} \in \mathcal{S}_{m,m}, |\tilde{q}_{ij} - \pi_j| \leq \beta\}.$$

Note that, for L large enough, $\mathbf{Q}^L \in \mathcal{A}(b(L)|\lambda_2(\mathbf{Q})|^L)$ for some polynomial $b(L)$. We will use $\beta \triangleq b(L)|\lambda_2(\mathbf{Q})|^L$ in the following. With this

$$\begin{aligned} C(\mathbf{Q}^L) &\leq \max_{\tilde{\mathbf{Q}} \in \mathcal{A}(\beta)} C(\tilde{\mathbf{Q}}) \\ &= \max_{\tilde{\mathbf{Q}} \in \mathcal{A}(\beta)} \max_{\mathbf{p}} I(\mathbf{p}, \tilde{\mathbf{Q}}) \\ &= \max_{\tilde{\mathbf{Q}} \in \mathcal{A}(\beta)} \min_{\mathbf{p}_Y} \max_{\mathbf{x} \in \mathcal{X}} D(\tilde{\mathbf{Q}}(\cdot|\mathbf{x})\|\mathbf{p}_Y) \\ &\leq \max_{\tilde{\mathbf{Q}} \in \mathcal{A}(\beta)} \max_{\mathbf{x} \in \mathcal{X}} D(\tilde{\mathbf{Q}}(\cdot|\mathbf{x})\|\boldsymbol{\pi}) \\ &= \max_{\tilde{\mathbf{Q}} \in \mathcal{A}(\beta)} D(\tilde{\mathbf{Q}}(\cdot|\mathbf{0})\|\boldsymbol{\pi}), \end{aligned}$$

where we have used the minimax formula for capacity [28] (Problem 2.3.1, page 147) and the symmetry of the set $\mathcal{A}(\beta)$.

Call $\tilde{\mathbf{q}}_0 \triangleq \tilde{\mathbf{Q}}(\cdot|\mathbf{0})$. The last maximization over all $\tilde{\mathbf{Q}} \in \mathcal{A}(\beta)$ can equivalently be expressed as maximization over all $\tilde{\mathbf{q}}_0 = \boldsymbol{\pi} + \mathbf{a}$ with $|a_i| \leq \beta$ and assuming $\beta \leq \pi_i$ for all i . By the convexity of divergence in both of its arguments, the maximizing $\tilde{\mathbf{q}}_0$ must be an extreme point of the set of admissible distributions. Hence $a_i \in \{\pm\beta, 0\}$ with at most one $a_i = 0$ and all $a_i \neq 0$ if m is even. Call \mathcal{I}^- the set of indices i such that $a_i = -\beta$ and \mathcal{I}^+ the set of i such that $a_i = \beta$. Note that $|\mathcal{I}^-| = |\mathcal{I}^+|$. With this, we can continue the above chain of inequalities

$$\begin{aligned} C(\mathbf{Q}^L) &\leq \sum_{i \in \mathcal{I}^+} (\pi_i + \beta) \log \left(1 + \frac{\beta}{\pi_i} \right) + \sum_{i \in \mathcal{I}^-} (\pi_i - \beta) \log \left(1 - \frac{\beta}{\pi_i} \right) \\ &\leq \sum_{i \in \mathcal{I}^+} (\pi_i + \beta) \frac{\beta}{\pi_i} - \sum_{i \in \mathcal{I}^-} (\pi_i - \beta) \frac{\beta}{\pi_i} \\ &\leq \beta^2 \sum_i \frac{1}{\pi_i}, \end{aligned}$$

where we have used the inequality $\log(1+x) \leq x$. Call $A \triangleq \sum_i \frac{1}{\pi_i}$. As \mathbf{Q} is primitive, all components of $\boldsymbol{\pi}$ are strictly positive. Thus $A < \infty$ and we get

$$\begin{aligned} \liminf_{L \rightarrow \infty} -\frac{1}{L} \log C(\mathbf{Q}^L) &\geq \liminf_{L \rightarrow \infty} -\frac{1}{L} \log ((b(L)|\lambda_2(\mathbf{Q})|^L)^2 A) \\ &= -2 \log |\lambda_2(\mathbf{Q})|. \end{aligned}$$

We will now show that $E_L(\mathbf{Q}) \leq -2 \log |\lambda_2(\mathbf{Q})|$. As before, we will use Lemma IV.2. Pick an arbitrary row of \mathbf{Q}^L , say $q^{(L)}(\cdot|\mathbf{x}_0)$. By (38) we can write for L large enough

$$q^{(L)}(\mathbf{y}|\mathbf{x}_0) = \pi(\mathbf{y}) + \beta_{\mathbf{y}},$$

with $\sum_{\mathbf{y}} \beta_{\mathbf{y}} = 0$ since the matrix is stochastic, and

$$\beta_{\mathbf{y}} \triangleq b_{\mathbf{y}}(L) |\lambda_2(\mathbf{Q})|^L \tag{39}$$

for some polynomials $b_{\mathbf{y}}(L)$. Using the fact that for the stationary distribution of a Markov chain $\pi \mathbf{Q}^L = \pi$ we have

$$\begin{aligned} C(\mathbf{Q}^L) &= \max_{\mathbf{p}} I(\mathbf{p}, \mathbf{Q}^L) \\ &\geq I(\pi, \mathbf{Q}^L) \\ &= \sum_{\mathbf{x}, \mathbf{y}} \pi(\mathbf{x}) q^{(L)}(\mathbf{y}|\mathbf{x}) \log \frac{q^{(L)}(\mathbf{y}|\mathbf{x})}{\sum_{\tilde{\mathbf{x}}} \pi(\tilde{\mathbf{x}}) q^{(L)}(\mathbf{y}|\tilde{\mathbf{x}})} \\ &= \sum_{\mathbf{x}} \pi(\mathbf{x}) \sum_{\mathbf{y}} q^{(L)}(\mathbf{y}|\mathbf{x}) \log \frac{q^{(L)}(\mathbf{y}|\mathbf{x})}{\pi(\mathbf{y})}. \end{aligned}$$

As

$$\sum_{\mathbf{y}} q^{(L)}(\mathbf{y}|\mathbf{x}) \log \frac{q^{(L)}(\mathbf{y}|\mathbf{x})}{\pi(\mathbf{y})}$$

is a divergence and hence nonnegative for every \mathbf{x} , we can further lower bound $C(\mathbf{Q}^L)$ by

$$\begin{aligned} C(\mathbf{Q}^L) &\geq \sum_{\mathbf{y}} \pi(\mathbf{x}_0) (\pi(\mathbf{y}) + \beta_{\mathbf{y}}) \log \left(1 + \frac{\beta_{\mathbf{y}}}{\pi(\mathbf{y})} \right) \\ &\geq \sum_{\mathbf{y}} \pi(\mathbf{x}_0) (\pi(\mathbf{y}) + \beta_{\mathbf{y}}) \left(\frac{\beta_{\mathbf{y}}}{\pi(\mathbf{y})} - \frac{3\beta_{\mathbf{y}}^2}{4\pi(\mathbf{y})^2} \right) \\ &= \sum_{\mathbf{y}} \frac{\pi(\mathbf{x}_0)}{\pi(\mathbf{y})} \left(\frac{1}{4}\beta_{\mathbf{y}}^2 - \frac{3}{4\pi(\mathbf{y})}\beta_{\mathbf{y}}^3 \right), \end{aligned}$$

where we have used the inequality $\ln(1+x) \geq x - \frac{1}{2}x^2 \geq x - \frac{3}{4}x^2$ in the second last step, which is valid for all $x \geq -\varepsilon$ for some $\varepsilon > 0$ and the fact that $\sum_{\mathbf{y}} \beta_{\mathbf{y}} = 0$. As \mathbf{Q} is primitive, all $\pi(\mathbf{y})$ are strictly positive and hence, for L large enough, $|\beta_{\mathbf{y}}/\pi(\mathbf{y})| \leq \varepsilon$ for any $\varepsilon > 0$. As the $\beta_{\mathbf{y}}^2$ term will dominate the $\beta_{\mathbf{y}}^3$ term for large L , we get using (39)

$$\liminf_{L \rightarrow \infty} -\frac{1}{L} \log C(\mathbf{Q}^L) \leq -2 \log |\lambda_2(\mathbf{Q})|.$$

■

Finally, the next lemma proves the upper bound in Theorem IV.5. Together with the other lemmas this concludes the proof of Theorem IV.5.

Appendix Lemma I.3. *Let \mathbf{Q} be a stochastic matrix and call $\tilde{\mathbf{Q}}$ the stochastic matrix obtained by deleting all inessential indices from \mathbf{Q} . Then*

$$E_L(\mathbf{Q}) \leq -2 \log |\lambda_2(\tilde{\mathbf{Q}})|,$$

with equality if $\mathbf{Q} = \tilde{\mathbf{Q}}$.

Proof: Call $\{\tilde{\mathbf{P}}_i\}$ the diagonal irreducible submatrices of $\tilde{\mathbf{Q}}$ and let $\{d_i\}$ be their corresponding periods. Define d to be the least common multiple of the $\{d_i\}$. Recall that $D(\mathbf{Q}) = D(\tilde{\mathbf{Q}}) = \sum_i d_i$ is the number of eigenvalues of modulus one of matrix \mathbf{Q} , and that, from Theorem IV.4, $\lim_{L \rightarrow \infty} C(\mathbf{Q}^L) = \log D(\mathbf{Q})$. For simplicity of notation we will use $D \triangleq D(\mathbf{Q})$ in the following.

Construct $\hat{\mathbf{Q}} \triangleq \tilde{\mathbf{Q}}^d$ and call $\{\hat{\mathbf{P}}_i\}_{i=1}^D$ its primitive diagonal submatrices. By (15), we can consider the powers of $\hat{\mathbf{Q}}$ instead of $\tilde{\mathbf{Q}}$. We will first lower bound $C(\mathbf{Q}^L)$ by restricting the support of the input distribution to contain only essential indices of \mathbf{Q} . Equivalently, we reduce \mathbf{Q} to the stochastic matrix $\tilde{\mathbf{Q}}$, containing only the essential indices of \mathbf{Q} . Now $\hat{\mathbf{Q}}$ is a block diagonal stochastic matrix for which, as it is a sum channel [17], we can compute capacity in terms of the capacity of its diagonal blocks $\{\hat{\mathbf{P}}_i\}_{i=1}^D$. From this, we have

$$C(\mathbf{Q}^{dL}) \geq C(\tilde{\mathbf{Q}}^{dL}) = \log \sum_{i=1}^D \exp C(\hat{\mathbf{P}}_i^L).$$

Hence

$$\begin{aligned} E_L(\mathbf{Q}) &= \liminf_{L \rightarrow \infty} -\frac{1}{dL} \log (C(\mathbf{Q}^{dL}) - \log(D)) \\ &\leq \liminf_{L \rightarrow \infty} -\frac{1}{dL} \log \left(\log \left(\sum_{i=1}^D \exp C(\hat{\mathbf{P}}_i^L) \right) - \log(D) \right) \\ &= \liminf_{L \rightarrow \infty} -\frac{1}{dL} \log \left(\log \left(\frac{1}{D} \sum_{i=1}^D \exp C(\hat{\mathbf{P}}_i^L) \right) \right), \end{aligned} \quad (40)$$

and we know from Appendix Lemma I.2 that

$$C(\hat{\mathbf{P}}_i^L) = b_i(L) \exp (2L \log |\lambda_2(\hat{\mathbf{P}}_i)|)$$

for some polynomial $b_i(L) \geq 0$.

We will now prove an auxiliary result, showing that for any $a_i \geq 0$ and polynomials $b_i(L) \geq 0$ we have

$$\lim_{L \rightarrow \infty} -\frac{1}{L} \log \log \left(\frac{1}{D} \sum_{i=1}^D \exp (b_i(L) \exp(-La_i)) \right) = \min_i a_i. \quad (41)$$

We will first show that the left hand side of (41) is upper bounded by $a_j \triangleq \min_i a_i$. As the arithmetic mean is greater than or equal to the geometric mean, we have

$$\begin{aligned}
& \lim_{L \rightarrow \infty} -\frac{1}{L} \log \log \left(\frac{1}{D} \sum_{i=1}^D \exp(b_i(L) \exp(-La_i)) \right) \\
& \leq \lim_{L \rightarrow \infty} -\frac{1}{L} \log \log \left(\prod_{i=1}^D \exp\left(\frac{1}{D} b_i(L) \exp(-La_i)\right) \right) \\
& = \lim_{L \rightarrow \infty} -\frac{1}{L} \log \left(\sum_{i=1}^D \frac{1}{D} b_i(L) \exp(-La_i) \right) \\
& \leq \lim_{L \rightarrow \infty} -\frac{1}{L} \log \left(\frac{1}{D} b_j(L) \exp(-La_j) \right) \\
& = a_j.
\end{aligned}$$

We will now show inequality in the other direction. For x positive and small enough there exists a $\varepsilon > 0$ such that $\exp(x) - 1 \leq x(1 + \varepsilon)$. Also, $\log(1 + x) \leq x$. From this, we have for L large enough

$$\begin{aligned}
& \lim_{L \rightarrow \infty} -\frac{1}{L} \log \log \left(\frac{1}{D} \sum_{i=1}^D \exp(b_i(L) \exp(-La_i)) \right) \\
& = \lim_{L \rightarrow \infty} -\frac{1}{L} \log \log \left(1 + \frac{1}{D} \sum_{i=1}^D \left(\exp(b_i(L) \exp(-La_i)) - 1 \right) \right) \\
& \geq \lim_{L \rightarrow \infty} -\frac{1}{L} \log \log \left(1 + \frac{1 + \varepsilon}{D} \sum_{i=1}^D b_i(L) \exp(-La_i) \right) \\
& \geq \lim_{L \rightarrow \infty} -\frac{1}{L} \log \left(\frac{1 + \varepsilon}{D} \sum_{i=1}^D b_i(L) \exp(-La_i) \right) \\
& \geq \lim_{L \rightarrow \infty} -\frac{1}{L} \log \left((1 + \varepsilon) b_j(L) \exp(-La_j) \right) \\
& = a_j.
\end{aligned}$$

Together, this proves the auxiliary result.

Assume that $|\lambda_2(\widehat{\mathbf{P}}_j)| \triangleq \max_i |\lambda_2(\widehat{\mathbf{P}}_i)|$ for all i . Then using the auxiliary result we get from (40) and (41)

$$\begin{aligned} E_L(\mathbf{Q}) &\leq -\frac{2}{d} \log |\lambda_2(\widehat{\mathbf{P}}_j)| \\ &= -2 \log |\lambda_2(\widehat{\mathbf{P}}_j)^{1/d}| \\ &= -2 \log |\lambda_2(\widetilde{\mathbf{Q}})|, \end{aligned}$$

where the last equality holds as the eigenvalues of $\widehat{\mathbf{Q}}$ are the the eigenvalues of $\widetilde{\mathbf{Q}}$ raised to the power d and as the eigenvalues of a block diagonal matrix are the union of the eigenvalues of its diagonal blocks. Moreover, we have equality in (40) (the only inequality used in the derivation) if $\mathbf{Q} = \widetilde{\mathbf{Q}}$. \blacksquare

APPENDIX II

PROOF OF LEMMA II.1

In this lemma, we derives conditions for a channel \mathbf{W} to be the worst under certain constraints. More specifi cally, consider a set Ω of (\mathbf{x}, \mathbf{y}) pairs and defi ne

$$\mathcal{W} \triangleq \left\{ \mathbf{W} : \mathbf{W} \in \mathcal{S}_{n,n}, w(\mathbf{y}|\mathbf{x}) \geq a(\mathbf{x}, \mathbf{y}), \sum_{(\mathbf{x}, \mathbf{y}) \in \Omega} (w(\mathbf{y}|\mathbf{x}) - a(\mathbf{x}, \mathbf{y})) \geq b \right\}$$

for some fi xed function $a(\mathbf{x}, \mathbf{y})$ and fi xed constant b .

Appendix Lemma II.1. *The pair (\mathbf{p}, \mathbf{W}) is the solution to the minimax problem*

$$\min_{\mathbf{W} \in \mathcal{W}} \max_{\mathbf{p}} I(\mathbf{p}, \mathbf{W})$$

if and only if \mathbf{p} and \mathbf{W} are valid (i.e., \mathbf{p} is a distribution on \mathcal{X} and $\mathbf{W} \in \mathcal{W}$) and if they satisfy the following conditions:

$$p(\boldsymbol{\alpha}) \log \frac{w(\boldsymbol{\gamma}|\boldsymbol{\alpha})}{\sum_{\tilde{\boldsymbol{x}}} p(\tilde{\boldsymbol{x}}) w(\boldsymbol{\gamma}|\tilde{\boldsymbol{x}})} \begin{cases} = \mu(\boldsymbol{\alpha}) + 1_{\Omega}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \lambda_b, & \text{if } w(\boldsymbol{\gamma}|\boldsymbol{\alpha}) > a(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \\ \geq \mu(\boldsymbol{\alpha}) + 1_{\Omega}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \lambda_b, & \text{if } w(\boldsymbol{\gamma}|\boldsymbol{\alpha}) = a(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \end{cases} \quad (42)$$

for all $\boldsymbol{\alpha} \in \mathcal{X}$, $\boldsymbol{\gamma} \in \mathcal{Y}$, for some vector $\boldsymbol{\mu}$ and for some $\lambda_b \geq 0$ such that

$$\left(\sum_{(\mathbf{x}, \mathbf{y}) \in \Omega} (w(\mathbf{y}|\mathbf{x}) - a(\mathbf{x}, \mathbf{y})) - b \right) \lambda_b = 0 \quad (43)$$

and

$$\sum_{\mathbf{y}} w(\mathbf{y}|\mathbf{x}) \log \frac{w(\mathbf{y}|\mathbf{x})}{\sum_{\tilde{\boldsymbol{x}}} p(\tilde{\boldsymbol{x}}) w(\mathbf{y}|\tilde{\boldsymbol{x}})} \begin{cases} = C, & \text{if } p(\mathbf{x}) > 0 \\ \leq C, & \text{if } p(\mathbf{x}) = 0 \end{cases} \quad (44)$$

for some $C \geq 0$.

Proof: The set of matrices \mathcal{W} and the set of possible distributions \mathbf{p} are both defined by linear inequalities and equalities and are hence convex sets. This implies, together with the fact that mutual information is convex in the channel transition probability matrix and concave in the input distribution, that

$$\min_{\mathbf{W} \in \mathcal{W}} \max_{\mathbf{p}} I(\mathbf{p}, \mathbf{W}) = \max_{\mathbf{p}} \min_{\mathbf{W} \in \mathcal{W}} I(\mathbf{p}, \mathbf{W}).$$

Moreover, for every fixed \mathbf{p} ,

$$\min_{\mathbf{W} \in \mathcal{W}} I(\mathbf{p}, \mathbf{W})$$

is a convex minimization problem in \mathbf{W} , and the Kuhn-Tucker conditions are necessary and sufficient for optimality [29]. The Lagrangian of the minimization problem is given by

$$\begin{aligned} \sum_{\mathbf{x}, \mathbf{y}} p(\mathbf{x}) w(\mathbf{y}|\mathbf{x}) \log \frac{w(\mathbf{y}|\mathbf{x})}{\sum_{\tilde{\mathbf{x}}} p(\tilde{\mathbf{x}}) w(\mathbf{y}|\tilde{\mathbf{x}})} - \sum_{\mathbf{x}, \mathbf{y}} (w(\mathbf{y}|\mathbf{x}) - a(\mathbf{x}, \mathbf{y})) \lambda_a(\mathbf{x}, \mathbf{y}) \\ - \left(\sum_{(\mathbf{x}, \mathbf{y}) \in \Omega} (w(\mathbf{y}|\mathbf{x}) - a(\mathbf{x}, \mathbf{y})) - b \right) \lambda_b - \sum_{\mathbf{x}, \mathbf{y}} \left(w(\mathbf{y}|\mathbf{x}) - \frac{1}{|\mathcal{X}|} \right) \mu(\mathbf{x}). \end{aligned} \quad (45)$$

To compute the Kuhn-Tucker conditions, we will have to find the derivative of (45) with respect to \mathbf{W} , which can be found to be

$$p(\boldsymbol{\alpha}) \log \frac{w(\boldsymbol{\gamma}|\boldsymbol{\alpha})}{\sum_{\tilde{\mathbf{x}}} p(\tilde{\mathbf{x}}) w(\boldsymbol{\gamma}|\tilde{\mathbf{x}})} - \lambda_a(\boldsymbol{\alpha}, \boldsymbol{\gamma}) - 1_{\Omega}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \lambda_b - \mu(\boldsymbol{\alpha}). \quad (46)$$

The Kuhn-Tucker conditions for optimality require then that (46) is equal to zero for all $\boldsymbol{\alpha}, \boldsymbol{\gamma}$ and for $\lambda_a(\boldsymbol{\alpha}, \boldsymbol{\gamma}), \lambda_b \geq 0$ together with the constraint that $\mathbf{W} \in \mathcal{W}$. Moreover, we have complementary slackness, that is

$$(w(\boldsymbol{\gamma}|\boldsymbol{\alpha}) - a(\boldsymbol{\alpha}, \boldsymbol{\gamma})) \lambda_a(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = 0 \quad (47)$$

for all $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ and (43). As all the $\lambda_a(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ are nonnegative, (46) and (47) can be combined to yield (42).

Similarly, we can derive Kuhn-Tucker conditions for the optimal \mathbf{p} for a fixed \mathbf{W} , as is done for example in [17], to obtain (44). Together, (42) and (44) yield a set of necessary and sufficient conditions for the optimal (\mathbf{p}, \mathbf{W}) pair. ■

Example II.1. Consider the set $\mathcal{W} \subset \mathcal{S}_{n,n}$ of stochastic matrices \mathbf{W} with $\text{tr}(\mathbf{W}) \geq nb$.

If $b > 1/n$ then

$$\mathbf{W}^* = \frac{1-b}{n-1} \mathbf{1}\mathbf{1}^T + \frac{nb-1}{n-1} \mathbf{I}$$

is an element of \mathcal{W} and together with $\mathbf{p}^* = \frac{1}{n} \mathbf{1}^T$ the pair $(\mathbf{p}^*, \mathbf{W}^*)$ satisfies the Kuhn-Tucker condition given in Appendix Lemma II.1. Hence \mathbf{W}^* is the worst channel in the set \mathcal{W} and the corresponding optimal input distribution is uniform.

If $b \leq 1/n$ then the worst channel is

$$\mathbf{W}^* = \frac{1}{n} \mathbf{1}\mathbf{1}^T,$$

and the Kuhn-Tucker conditions are satisfied for the pair $(\mathbf{p}, \mathbf{W}^*)$ for any choice of \mathbf{p} . In this case, the resulting capacity is always zero. \diamond

APPENDIX III

CONNECTION BETWEEN THE ZERO ERROR CAPACITY AND COMMON INFORMATION

An intuitive interpretation of the results about infinitely long cascade of channels can be obtained by introducing the concept of *common information* as defined in [24]. Consider random variables X and Y with \mathbf{p} as marginal distribution of X and \mathbf{V} as conditional distribution. Assume, for the moment, that the marginal distributions of X and Y are strictly positive. Partition the input alphabet into disjoint sets $\mathcal{X} = \bigcup_i \mathcal{X}_i$ such that x and x' are in the same \mathcal{X}_i if there exists a sequence $a_1, b_1, a_2, \dots, b_{n-1}, a_n$ with $a_j \in \mathcal{X}$, $b_j \in \mathcal{Y}$, $a_1 = x$, $a_n = x'$ such that

$$\prod_{j=1}^{n-1} \Pr[Y = b_j | X = a_j] \Pr[X = a_{j+1} | Y = b_j] > 0.$$

Similarly, we partition $\mathcal{Y} = \bigcup_i \mathcal{Y}_i$. It is easily seen that from any given \mathcal{X}_i there is only one \mathcal{Y}_j such that

$$\Pr[X \in \mathcal{X}_i | Y \in \mathcal{Y}_j] = \Pr[Y \in \mathcal{Y}_j | X \in \mathcal{X}_i] = 1,$$

and we can always relabel one of the partitions such that this will be the case when $i = j$. This is equivalent to saying that there exist two permutation matrices $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ such that

$$\mathbf{\Pi}_1 \text{diag}(\mathbf{p}) \mathbf{V} \mathbf{\Pi}_2 = \begin{pmatrix} \mathbf{P}_1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & & \ddots & \vdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_D & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad (48)$$

where D is the cardinality of $\{\mathcal{X}_i\}$ (or, equivalently, $\{\mathcal{Y}_i\}$). Note that we have a zero block on the diagonal of (48), which accounts for elements of \mathcal{X} and \mathcal{Y} occurring with probability zero (i.e., for which the marginal distribution is zero). Define a random variable S taking values on $\{1, \dots, D\}$ with

$$\Pr[S = i] \triangleq \Pr[X \in \mathcal{X}_i] = \Pr[Y \in \mathcal{Y}_i].$$

The common information between X and Y is then defined to be [24]

$$J(\mathbf{p}, \mathbf{V}) \triangleq H(S).$$

We will now derive a connection between the common information and the zero error capacity of a channel.

Appendix Proposition III.1.

$$\max_{\mathbf{p}} J(\mathbf{p}, \mathbf{V}) = \log M_0(\mathbf{V}).$$

Remark. This result implies that

$$C_0(\mathbf{V}) = \sup_n \frac{1}{n} \max_{\mathbf{p}} J(\mathbf{p}, \mathbf{V}^{\otimes n}).$$

Hence zero error capacity can be defined through common information in a similar manner as ordinary capacity can be defined through mutual information. The crucial difference is, however, that for ordinary capacity we have a single letter characterization, whereas for zero error capacity we have to take a supremum over all n .

Proof: From any triple \mathbf{p}, Π_1, Π_2 as in the definition of common information we can construct an encoder M_E and a decoder M_D which define a zero error code of rate $\log(D) \geq H(S)$. Hence

$$\max_{\mathbf{p}} J(\mathbf{p}, \mathbf{V}) \leq \log M_0(\mathbf{V}).$$

To show inequality in the other direction, note that for every encoder M_E defining a zero error code with rate $\log(D)$ there exists a \mathbf{p} assigning all codewords of M_E the same positive probability and all other values in \mathcal{X} probability zero. As the codewords of M_E are non adjacent, they all lie in different \mathcal{X}_i . For this \mathbf{p} , we have $H(S) = \log(D)$ and thus

$$\max_{\mathbf{p}} J(\mathbf{p}, \mathbf{V}) \geq \log M_0(\mathbf{V}).$$

As

$$\begin{aligned}
 I(\mathbf{p}, \mathbf{V}) &= H(X) - H(X|Y) \\
 &= H(X, S) - H(X|Y, S) \\
 &\geq H(X|S) + H(S) - H(X|S) \\
 &= J(\mathbf{p}, \mathbf{V}),
 \end{aligned}$$

we see that mutual information is always greater than or equal to common information. Appendix Proposition III.1 and Theorem V.3 show, however, that when infinitely many (identical) channels are to be cascaded, the two are the same. Hence as $L \rightarrow \infty$, the only part of mutual information we can preserve between the input and the output of the cascade is exactly the common information between them.

REFERENCES

- [1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.
- [2] R. Ahlswede, N. Cai, S-Y. R. Li, and R. W. Yeung, "Network information flow," pp. 1204–1216, July 2000.
- [3] S-Y. R. Li, R. W. Yeung, and N. Cai, "Linear network coding," vol. 49, pp. 371–381, Feb. 2003.
- [4] P. Sanders, S. Egner, and L. Tolhuizen, "Polynomial time algorithms for network information flow," *Proc. 15th ACM Symposium on Parallel Algorithms and Architectures*, 2003.
- [5] R. Koetter and M. Médard, "Beyond routing: an algebraic approach to network coding," *IFOCOM*, vol. 1, pp. 122–130, June 2002.
- [6] C. Fragouli and E. Soljanin, "Information flow decomposition for network coding."
- [7] C. E. Shannon, "The zero error capacity of a noisy channel," *IEEE Transactions on Information Theory*, vol. 2, no. 3, pp. 8–19, September 1956.
- [8] P. Gupta and P. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, pp. 388–404, Mars 2000.
- [9] D. Tuninetti and C. Fragouli, "Processing along the way: Forwarding vs. coding," in *Proceedings of the International Symposium on Information Theory and its Applications*, Parma, Italy, October 2004.
- [10] T. C. et al., "Presentation at private workshop after isita 2004," October 2004.
- [11] M. K. Simon, "On the capacity of a cascade of identical discrete memoryless nonsingular channels," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 100–102, January 1970.
- [12] R. A. Silverman, "On binary channels and their cascade," *IEEE Transactions on Information Theory*, vol. 1, no. 3, pp. 19–27, December 1955.
- [13] A. B. Kiely and J. T. Coffey, "On the capacity of a cascade of channels," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1310–1321, July 1993.

- [14] D. Lun, M. Médard, and M. Effros, "On coding for reliable communication over packet networks," *Proc. 412st Annual Allerton Conference*, Monticello, IL, Oct. 2004.
- [15] P. Pakzad, C. Fragouli, and A. Shokrollahi, "Coding schemes for line networks," *ISIT*, 2005.
- [16] J. Körner and A. Orłitsky, "Zero-error information theory," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2207–2229, October 1998.
- [17] R. G. Gallager, *Information Theory and Reliable Communication*. Wiley, 1968.
- [18] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1996.
- [19] E. Seneta, *Non-negative Matrices and Markov Chains*. Springer Verlag, 1981.
- [20] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.
- [21] —, *Topics in Matrix Analysis*. Cambridge University Press, 1991.
- [22] C. E. Shannon, "Some geometrical results in channel capacity," *Nachrichtentechnische Zeitschrift*, vol. 10, 1957.
- [23] M. Saks and A. Condon, "A limit theorem for sets of stochastic matrices," *Linear Algebra and Applications*, vol. 381, pp. 61–76, 2004.
- [24] P. Gács and J. Körner, "Common information is far less than mutual information," *Problems of Control and Information Theory*, vol. 2, no. 2, pp. 149–162, 1973.
- [25] L. Lovász, "On the Shannon capacity of a graph," *IEEE Transactions on Information Theory*, vol. 25, no. 1, pp. 1–7, January 1979.
- [26] V. Vazirani, *Approximation Algorithms*. Springer, 2001.
- [27] M. Chiang and S. Boyd, "Geometric programming duals of channel capacity and rate distortion," *IEEE Transactions on Information Theory*, vol. 50, no. 2, pp. 245–258, February 2004.
- [28] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Akadémiai Kiadó, 1981.
- [29] S. Boyd and L. Vanderberghe, *Convex Optimization*. Cambridge University Press, 2004.